III. Equations

I.P.M. Solution of equations according to the criteria enunciated in the Teaching Remarks below. For example,

$$2x^2 - 5x - 4 = 0,$$

the coefficients being exact. Then

$$x = \frac{5 \pm \sqrt{25 + 32}}{4},$$

$$x \approx \frac{5 \pm 7.55}{4},$$

and

$$x_1 \approx 3.14, \quad x_2 \approx -0.64.$$

T.R. The solution of equations leads very often to approximate roots. This may happen in the case of quadratic equations as well as in the case of simultaneous linear equations. In either of the cases, if the coefficients of the unknowns are approximate, the roots of the equations are invariably approximate. If the coefficients of the unknowns are exact, but at least one root is approximate, all the remaining roots are invariably approximate. The approximativeness of the roots of the equations, if the coefficients of the unknowns are approximate, depends on the approximativeness of these coefficients. If the coefficients of the unknowns are exact, and one of the roots, at least, is approximate, the approximativeness of all other roots (generally precision) should conform to the determined approximativeness of the first obtained root.

CHAPTER X

APPROXIMATE COMPUTATION: APPROXIMATE COMPUTATION OF THE SECOND TYPE

Approximate Computation of the second type, that is, the process of computation in which either the approximativeness of the final numerical result is predetermined, or the approximativeness of a measurement is to be determined so that a given numerical result shall be of a given approximativeness, is to a certain extent more difficult than Approximate Computation of the first type. A few selected problems involving Approximate Computation of the second type will give the pupils an insight into the processes that are very common in scientific experiments involving measurement, and some problems may even represent processes that are common in trades and engineering work.

I. The Notion of a Predetermined Approximativeness

I.P.M. Illustrative material showing that in many cases the approximativeness of numerical data with which certain numerical processes are to be performed should conform with the predetermined approximativeness of the final numerical result (see section II below). In many other cases approximate data (or measurement data) are not secured yet, but the predetermined approximativeness of the final numerical result is given. In such a case the question that arises is: "How well should the measurement be performed?" (See section III below.)

T.R. The teacher should select type-problems illustrating the fact that in many cases it is necessary to perform measurements in such a manner that a desired approximativeness is obtained. For example; in measuring areas it is at times required to obtain the measure of a given area so that this measure does not exceed a certain error; that is, the measure should be correct to a certain unit. In order to avoid unnecessary guessing, the teacher should show that it is possible to compute the precision of the linear measurement. This process represents Approximate Computation of the second type which should have a place in secondary school mathematics.

II. DETERMINATION OF APPROXIMATIVENESS OF DATA IN ORDER TO COMPLY WITH LIMITATIONS OF APPROXIMATIVENESS OF THE RESULT SET FORTH IN THE PROBLEM

I.P.M. Examples of the type: The sides of a rectangle are 64.53 and 472.47 inches each. It is required to obtain the area of the rectangle correct to hundreds of square inches. We observe that 64 times 472 will give about 30,000. Then the final numerical result is expected to be correct to three significant digits. We round off the two approximate numbers as follows: 64.5 and 472.5. The final numerical result is then 30,476.25, and correctly stated as required by the problem, we have 30,500. square inches.

T.R. At times the numerical data exceed in their approximativeness the information necessary for approximativeness of the final numerical result. If so, the teacher should show the pupils that it is absurd to perform the numerical work with actually unnecessary digits. It is logical and consistent then to cut down the number of significant digits so that unnecessary numerical work is eliminated. Naturally, this should be performed in accordance with the rules for multiplication and (or) division stated in Chapter VI.

III. DETERMINATION OF APPROXIMATIVENESS OF MEASUREMENTS TO BE PERFORMED IN ORDER TO OBTAIN A FINAL NUMERICAL RESULT OF PREDETERMINED APPROXIMATIVENESS

I.P.M. Examples of the type: The area of a square is about 1,500 square inches. It is required to measure the side of the square so that after the area is computed the error in the area does not exceed 5 square inches. Let the error in measurement of the side of the square be x. The side of the square is taken to be $\sqrt{1{,}500} = 38.73$. Then the error of the area of the square is

$$5 = 2 \cdot 38.73x,$$

and

$$x = \frac{5}{77.46} = 0.07 \text{ (approximately).}$$

In other words, the measurement of the side of the square should be correct to tenths of an inch.

T.R. When a measurement has not been performed, but the approximate value of the numerical result is known, and it is required to determine the approximativeness with which to perform certain measurements so that a definite approximativeness is obtained, recourse should be made to the formula for the apparent error of the formula by means of which this final result is computed after the required measurement or measurements are performed. The approximativeness of the required measurement is obtained by means of solving this expression in which the apparent error of the measurement appears as an unknown quantity.

This type of problem requires careful analysis and sound judgment. The teacher should bear in mind, however, that in practice such problems are very common both in the experience of the scientist and in the practical work of an engineer. In school work in physics the pupils are confronted with this type of problem. For example, the following problem belongs to this category: What should be the length of a seconds pendulum in order that the period of its vibration will be correct to 0.1, 0.01, or 0.001 of a second?

CHAPTER XI

APPROXIMATE COMPUTATION IN INTERMEDIATE ALGEBRA

1. GENERAL REMARKS

THE work in Approximate Computation in intermediate algebra should be a continuation of the work begun in elementary algebra and numerical trigonometry. All the concepts concerning Approximations and Approximate Computation introduced and developed during the study of elementary algebra and numerical trigonometry should be assumed thereafter to be known to the pupils. The aim of the work in Approximate Computation in intermediate algebra is the development of the notion of the influence of errors on the approximativeness of numerical results of performed computations as well as the gradual development of the technique of strict evaluation of this approximativeness in terms of apparent errors.

Beginning with intermediate algebra the numerical value of an apparent error should play an important part in the statement of a measurement as expressed by means of an approximate number. The pupils should learn from practice of the importance of this apparent error, and they should gradually develop the habit of stating the numerical value of a measurement as well as the numerical results of performed computations with their respective errors. The pupils should also develop the habit of insisting that the data with which they have to work shall be given with their respective errors. They should become convinced that an approximate number given without its apparent error (or relative error) is meaningless. The use of the so-called "greatest possible error," that is, the apparent error of the form ±0.5 of a unit of measure in terms of which a measurement is expressed, should be made with discretion. The teacher should try to vary the magnitude of apparent errors. Only when no means are available for obtaining the magnitude of an apparent error should the teacher allow the pupils to resort to the greatest possible error.

In the development of the rules for numerical operations with approximate numbers these rules will be derived in terms of formulas. The translation of these formulas into verbal language would result in long and abstruse statements. The teacher should point out to the pupils the power of an algebraic statement, its elegance, conciseness, succinctness, and directness. He should make a special effort to emphasize the fact that the symbolic language of algebra is more powerful than any other language.

II. Representation of Approximate Numbers

The work in Approximate Computation on the level of intermediate algebra should be based on the representation of approximate numbers, or measures, as being located between two measures known exactly. This representation should be tied up with the apparent error in the following manner: An approximate number, for example, $a = 42.3 \pm 0.04$, is represented as

$$42.26 < a < 42.34.$$

Conversely, if an approximate number is given in the form of an inequality, the pupils should be able to rewrite it in the equivalent form with the apparent error. For example,

$$64.78 < a < 64.82,$$

then

$$A = \frac{64.82 + 64.78}{2} = 64.8,$$

and

$$e = \frac{64.82 - 64.78}{2} = 0.02.$$

Finally,

$$a = 64.8 \pm 0.02.$$

The numbers 42.26 and 64.78 are known as the lower bounds and the numbers 42.34 and 64.82 are known as the upper bounds of their respective approximate numbers. Symbolically we have then

$$A_2 < a < A_1,$$

or

$$A_2 < A < A_1,$$

and

$$A = \frac{A_1 + A_2}{2}, \text{ and } e = \frac{A_1 - A_2}{2}.$$

III. Rounding Off Digits

In most of the numerical results that the pupils will obtain, the apparent errors will contain several significant digits. The pupils should round off approximate numbers so that apparent errors contain one significant digit only. Errors when rounded off should always be forced; see example below. In connection with the process of rounding off the teacher should emphasize the fact that the interval (A_2, A_1) in which a is located should never be narrowed, because this would be contrary to the nature of the process of measurement. We never can make a measure more precise, after it has been obtained. The interval (A_2, A_1) may be widened, and when rounding off takes place it should be widened. For example, $a = 25.537 \pm 0.042$ is rounded off as follows:

$$25.53 \pm 0.049, \quad (0.042 + 0.007 = 0.049),$$

or

$$25.53 \pm 0.05;^*$$

again

$$25.5 \pm 0.08, \quad (0.05 + 0.03 = 0.08).$$

The approximate number may be left in this last form, because it thus contains significant digits only, or if desired it can be rounded off further as follows:

$$25. \pm 0.58, \quad (0.08 + 0.5 = 0.58),$$

and finally, it should be written as $25. \pm 0.6$, whereby one significant digit is lost. Further rounding off would result in:

$$20 + 5.6, \quad \text{or} \quad 20 + 6, \quad \text{or} \quad 30 - 4$$

An ample number of illustrative examples should be used by the teacher in order to emphasize the importance of the apparent errors in the process of rounding off digits.

If an approximate number is given in the form of its bounds and it is required to round it off, no digit in the lower bound should ever be forced, whereas the digits of the upper bound should always be forced, even though the first digit to be dropped is one of the following: 0, 1, 2, 3, 4, or 5. This procedure is consistent with the fact that the interval (A_2, A_1) may only be widened. For example,

$$236{,}014 < a < 237{,}031$$

is rounded off as follows:

$$236{,}010 < a < 237{,}040,$$
$$236{,}000 < a < 237{,}100,$$
$$230{,}000 < a < 238{,}000.$$

---

[*] Some writers use this method of writing approximate numbers. But if this method is used, the last digit on the extreme right of the approximate number is not significant; it is doubtful.

IV. NUMERICAL PROCESSES

All numerical processes should be performed in accordance with the rules for operations with inequalities.

A. ADDITION

Let

$$A_2 < a < A_1$$

and

$$B_2 < b < B_1.$$

Then

$$A_2 + B_2 < a + b < A_1 + B_1.$$

For example

$$a = 16.4 \pm 0.04,$$

and

$$b = 14.3 \pm 0.02.$$

The numerical work should be performed in the following manner (L.B. denotes the lower bound, and U.B. denotes the upper bound):

	L.B.	U.B.
a	16.36	16.44
b	14.28	14.32
$x = a + b$	30.64	30.76

$$x = \frac{30.76 + 30.64}{2} = 30.4$$

$$e = \frac{30.76 - 30.64}{2} = 0.06$$

Finally

$$a + b = 30.4 \pm 0.06.$$

Suppose we have

$$a = 62.3 + 0.04,$$

and

$$b = 31.9 \pm 0.03.$$

We proceed as follows:

	L.B.	U.B.
a	62.30	62.34
b	31.87	31.93
$x = a + b$	94.17	94.27

$$x = \frac{94.27 + 94.17}{2} = 94.22$$

$$e = \frac{94.27 - 94.17}{2} = 0.05$$

Finally,

$$a + b = 94.22 \pm 0.05, \quad \text{or} \quad a + b = 94.2 \pm 0.07,$$

or, if we wish, we can write:

$$a + b = 94. + 0.3.$$

B. SUBTRACTION

Let

$$A_2 < a < A_1$$

and

$$B_2 < b < B_1.$$

Then

$$A_2 - B_1 < a - b < A_1 - B_2.$$

For example,

$$a \approx 62.3 \pm 0.04,$$

and

$$b = 31.9 \pm 0.03.$$

Then

	L.B.	U.B.
a	62.26	62.34
(!) b	31.93	31.87
$x = a - b$	30.33	30.47

*

$$x = \frac{30.47 + 30.33}{2} = 30.4$$

$$e = \frac{30.47 - 30.33}{2} = 0.07$$

Finally,

$$a - b \approx 30.4 \pm 0.07,$$

or, if we wish, we can write:

$$a - b \approx 30. + 0.5.$$

C. MULTIPLICATION

Let

$$A_2 < a < A_1$$

and

$$B_2 < b < B_1.$$

Then

$$A_2B_2 < ab < A_1B_1.$$

For example,

$$a = 43.7 \pm 0.03,$$

and

$$b = 54.6 \pm 0.04.$$

Then

	L.B.	U.B.
a	43.67	43.73
b	54.56	54.64
$x = ab$	2,382.6	2,389.4

$$x = \frac{2{,}389.4 + 2{,}382.6}{2} = 2{,}386$$

$$e = \frac{2{,}389.4 - 2{,}382.6}{2} = 3.4$$

Finally

$$ab \approx 2{,}386 \pm 4.$$

---

* The exclamation sign (!) denotes the fact that the bounds are reversed.

D. DIVISION

Let

$$A_2 < a < A_1$$

and

$$B_2 < b < B_1.$$

Then

$$\frac{A_2}{B_1} < \frac{a}{b} < \frac{A_1}{B_2}$$

For example,

$$a \approx 53.7 \pm 0.03,$$

and

$$b \approx 21.5 \pm 0.04.$$

Then

	L.B.	U.B.
a	53.67	53.73
(!) b	21.54	21.46
$x = a : b$	2.4916	2.5037

$$x = \frac{2.5037 + 2.4916}{2} \approx 2.4977$$

$$e = \frac{2.5037 - 2.4916}{2} \approx 0.0061$$

Finally,

$$a : b \approx 2.4977 \pm 0.0061,$$

or

$$a : b \approx 2.498 \pm 0.007.$$

E. RAISING TO A POWER

Let

$$A_2 < a < A_1,$$

then

$$A_2^2 < a^2 < A_1^2.$$

For example,

$$a \approx 42.7 \pm 0.04.$$

Then

	L.B.	U.B.
a	42.66	42.74
$x = a^2$	1,819.8	1,826.7

$$x = \frac{1{,}826.7 + 1{,}819.8}{2} \approx 1{,}823.3$$

$$e = \frac{1{,}826.7 - 1{,}819.8}{2} \approx 3.5$$

Finally

$$a^2 \approx 1{,}823.3 \pm 3.5,$$

or

$$a^2 \approx 1{,}823 \pm 4.$$

F. EXTRACTION OF ROOTS

Let

$$A_2 < a < A_1,$$

then

$$\sqrt{A_2} < \sqrt{a} < \sqrt{A_1}.$$

For example,

$$a = 51.3 \pm 0.03.$$

Then

	L.B.	U.B.
a	51.27	51.33
$x = \sqrt{a}$	7.1603	7.1645

$$x = \frac{7.1645 + 7.1603}{2} = 7.1628$$

$$e = \frac{7.1645 - 7.1603}{2} = 0.0021$$

Finally

$$\sqrt{a} = 7.1628 \pm 0.0021,$$

or

$$\sqrt{a} = 7.16 \quad \pm 0.005.$$

V. EVALUATION OF FORMULAS

Evaluation of formulas is performed as in section IV above.

Let, for example,

$$c = \frac{ab}{d},$$

where

$$a = 45.3 \pm 0.05,$$
$$b = 64.7 \pm 0.02,$$

and

$$d = 27.4 \pm 0.03.$$

Then

	L.B.	U.B.
a	45.25	45.35
b	64.68	64.72
ab	2,926.7	2,935.1
(!) d	27.43	27.37
$c = \frac{ab}{d}$	106.69	107.24

$$x = \frac{107.24 + 106.69}{2} = 106.97$$

$$e = \frac{107.24 - 106.70}{2} = 0.27$$

Finally,

$$c = 106.97 \pm 0.27, \quad \text{or} \quad c = 106.9 \pm 0.34,$$

or

$$c = 107. \pm 0.5.$$

As another illustration, consider the following problem. Calculate $\sin 15°$, if $\sin 45° = \cos 45° = \frac{1}{2}\sqrt{2}$, $\sin 30° = \frac{1}{2}$, and $\cos 30° = \frac{1}{2}\sqrt{3}$ and the angles are exact. We have then

$$\sin 15° = \sin (45° - 30°) = \sin 45° \cdot \cos 30° - \cos 45° \cdot \sin 30°,$$

or

$$\sin 15° = \tfrac{1}{4}(\sqrt{6} - \sqrt{2}).$$

Then

$$2.44948 < \sqrt{6} < 2.44950,$$
$$1.41420 < \sqrt{2} < 1.41422,$$

and

	L.B.	U.B.
$\sqrt{6}$	2.44948	2.44950
(!) $\sqrt{2}$	1.41422	1.41420
$4 \cdot \sin 15°$	1.03526	1.03530
$\sin 15°$	0.25881	0.25883

We then get that

$$\sin 15° = 0.2588 \pm 0.00003.$$

All other rules for Approximate Computation stated in the outline for elementary algebra and numerical trigonometry hold for the processes with bounds. As a rule, in all intermediate processes one doubtful (non-significant) digit should be retained, and the final result should be stated in terms of significant digits.

CHAPTER XII

APPROXIMATE COMPUTATION IN ADVANCED ALGEBRA AND TRIGONOMETRY

I. GENERAL REMARKS

THE work in Approximate Computation in advanced algebra and trigonometry, which represents a continuation of the work done in intermediate algebra, should also be concentrated on the evaluation of the approximativeness of numerical results by means of errors. In other words, the pupils should be given ample opportunity to make practical use of the concepts and techniques which were developed in their previous school work in order that they may be thoroughly prepared to master Approximate Computation in their college work in mathematics and science.

No new concepts need be developed during the work in Approximate Computation in advanced algebra and trigonometry. But the aim of the work at these levels should be the development of mastery of the subject, and ability to perform the numerical processes quickly and efficiently.

II. NUMERICAL PROCESSES

The work in Approximate Computation in advanced algebra and trigonometry should revert once more to the use of errors. At this stage, the pupils should, by means of bounds, be able to obtain the expressions for the apparent errors that arise from the different numerical processes. The following symbolism will be used in the theoretical development of this work:

The approximate number $a = A \pm e$,
then

$$A_2 < a < A_1, \quad \text{or} \quad A_2 < A < A_1,$$

and

$$A_2 = A - e, \quad \text{and} \quad A_1 = A + e,$$

where

$$A = \frac{A_1 + A_2}{2}, \quad \text{and} \quad e = \frac{A_1 - A_2}{2}.$$

A. ADDITION

Let

$$A_2 < a < A_1,$$

and

$$B_2 < b < B_1,$$

where

$$a = A \pm e_1, \quad \text{and} \quad b = B \pm e_2.$$

Then

$$A_2 + B_2 < a + b < A_1 + B_1,$$

and the apparent error of $(a + b)$ is

$$\frac{A_1 + B_1 - (A_2 + B_2)}{2} = \frac{A_1 - A_2}{2} + \frac{B_1 - B_2}{2},$$

or, it is

$$e_1 + e_2.$$

For example

$$\begin{array}{r} a = 162.7 \pm 0.03 \\ b = 179.5 \pm 0.04 \\ \hline a + b = 342.2 \pm 0.07, \end{array}$$

or this can be rounded off

$$a + b = 342. \quad \pm 0.3.$$

B. SUBTRACTION

Let

$$A_2 < a < A_1,$$

and

$$B_2 < b < B_1,$$

where

$$a = A \pm e_1, \quad \text{and} \quad b = B \pm e_2.$$

Then

$$A_2 - B_1 < a - b < A_1 - B_2,$$

and the apparent error of $(a - b)$ is

$$\frac{A_1 - B_2 - (A_2 - B_1)}{2} = \frac{A_1 - A_2}{2} + \frac{B_1 - B_2}{2},$$

or, it is

$$e_1 + e_2.$$

For example,

$$\begin{array}{r} a = 347.8 \pm 0.04 \\ b = 125.6 \pm 0.03 \\ \hline a - b = 222.2 \pm 0.07, \end{array}$$

or, the difference can be rounded off (if desired) to

$$a - b = 222. \quad \pm 0.3.$$

C. MULTIPLICATION

Let

$$A_2 < a < A_1,$$

and

$$B_2 < b < B_1,$$

where

$$a = A \pm e_1, \quad \text{and} \quad b = B \pm e_2.$$

Then

$$A_2B_2 < ab < A_1B_1.$$

This inequality can be rewritten as follows:

$$(A - e_1)(B - e_2) < ab < (A + e_1)(B + e_2).$$

The apparent error of the product ab is then

$$\frac{(A + e_1)(B + e_2) - (A - e_1)(B - e_2)}{2} = Ae_2 + Be_1.$$

For example,

$$\begin{array}{rl} a \approx & 24.3 \pm 0.02 \\ b \approx & 31.5 \pm 0.03 \\ \hline ab \approx & 765.5 \pm 1.359 \end{array}$$

As an alternative, the relative error may be used. The relative error of a product of two approximate numbers is

$$\frac{Ae_2 + Be_1}{AB} = \frac{e_1}{A} + \frac{e_2}{B}.$$

Then, we have

$$\begin{array}{rlr} a \approx & 24.3 \pm 0.02; & 0.00082 \\ b \approx & 31.5 \pm 0.03; & 0.00095 \\ \hline ab \approx & 765.5 & \pm 765.5 \cdot 0.00177 \end{array}$$

The apparent error of the product ab is 1.35.
The above result is then rounded off to

$$ab \approx 765 \pm 2.$$

D. DIVISION

Let

$$A_2 < a < A_1,$$

and

$$B_2 < b < B_1,$$

where

$$a = A \pm e_1, \quad \text{and} \quad b = B \pm e_2.$$

Then

$$\frac{A_2}{B_1} < \frac{a}{b} < \frac{A_1}{B_2}.$$

This inequality can be rewritten as follows:

$$\frac{A - e_1}{B + e_2} < \frac{a}{b} < \frac{A + e_1}{B - e_2}.$$

The apparent error of the quotient $a : b$ is then

$$\tfrac{1}{2}\left[\frac{A + e_1}{B - e_2} - \frac{A - e_1}{B + e_2}\right] \approx \frac{Ae_2 + Be_1}{B^2}.$$

Since the amount of numerical work involved in the calculation of the apparent error of a quotient makes its use prohibitive, we shall use the relative error of the quotient, which is

$$\frac{Ae_2 + Be_1}{B^2} : \frac{A}{B} = \frac{e_1}{A} + \frac{e_2}{B}.$$

For example,

$$\begin{array}{lll} a = 576.4 \pm 0.03; & & 0.000052 \\ b = 123.2 \pm 0.02; & & 0.000162 \\ \hline \dfrac{a}{b} = \quad 4.679 & \pm 4.679 \cdot & 0.000214. \end{array}$$

The apparent error of the quotient is 0.001.

Then

$$\frac{a}{b} = 4.679 \pm 0.001,$$

or

$$\frac{a}{b} = 4.68 \pm 0.002.$$

E. RAISING TO A POWER

Let

$$A_2 < a < A_1,$$

where

$$a = A \pm e.$$

Then

$$A_2^2 < a^2 < A_1^2.$$

This inequality can be rewritten as follows:

$$(A - e)^2 < a^2 < (A + e)^2.$$

The apparent error of the square is

$$\frac{(A + e)^2 - (A - e)^2}{2} = 2Ae,$$

and the relative error of the square is $2\dfrac{e}{A}$.

For example,

$$a \approx 64.7 \pm 0.03; \qquad 0.00046,$$
$$a^2 \approx 4,186. \qquad \pm 4,186 \cdot 0.00092.$$

The apparent error of the square is 3.
Then

$$a^2 \approx 4190 \pm 7.$$

The derivation of the expression for the apparent error of a^n may be effected by means of the binomial expansion of $(A \pm e)^n$. The reader will find no difficulty in obtaining the expression

$$nA^{n-1}e$$

for the apparent error, and the expression

$$n\frac{e}{A}$$

for the relative error of a^n.

F. EXTRACTION OF ROOTS

Let

$$A_2 < a < A_1,$$

where

$$a \approx A \pm e.^*$$

Then

$$\sqrt{A_2} < \sqrt{a} < \sqrt{A_1}.$$

This inequality can be rewritten as follows:

$$\sqrt{A - e} < \sqrt{a} < \sqrt{A + e}.$$

Let

$$x = \sqrt{A \pm e} = B \pm e_1, \text{ where } B = \sqrt{A}.$$

Then

$$A \pm e = B^2 \pm 2Be_1 + e_1^2 = A \pm 2\sqrt{A}\, e_1 + e_1^2.$$

Disregarding e_1^2, we have: $e \approx 2\sqrt{A}\, e_1$,
and

$$e_1 \approx \frac{e}{2\sqrt{A}}.$$

The inequality

$$\sqrt{A - e} < \sqrt{a} < \sqrt{A + e}$$

---

* It is assumed that $A_2 \geqslant 1$. In case $0 < A_2 < A_1 < 1$, we have a combination of division and extraction of a square root.

can then be rewritten as follows:

$$\sqrt{A} - \frac{e}{2\sqrt{A}} < \sqrt{a} < \sqrt{A} + \frac{e}{2\sqrt{A}},$$

and the apparent error of the square root is

$$\frac{e}{2\sqrt{A}},$$

while the relative error of the square root is

$$\tfrac{1}{2}\,\frac{e}{A}.$$

For example,

$$\begin{array}{rlr} a & \approx 7.17 \pm 0.002; & 0.00028, \\ \sqrt{a} & \approx 2.6777 & \pm 2.6777 \cdot 0.00014. \end{array}$$

The apparent error of the square root is 0.0004,
then

$$\sqrt{a} = 2.678 \pm 0.0007.$$

The derivation of the expression for the apparent error of $\sqrt[n]{a}$ may be effected by means of the binomial expansion of $(A \pm e)^{1/n}$. The reader will find no difficulty in obtaining the expression

$$\frac{e}{n\sqrt[n]{A^{n-1}}}$$

for the apparent error, and the expression

$$\frac{1}{n}\,\frac{e}{A}$$

for the relative error of $\sqrt[n]{a}$.

III. LOGARITHMIC FUNCTIONS

The work with logarithms of natural numbers and of trigonometric functions should preferably involve the use of apparent errors of these functions. In other words, this procedure would necessitate the tabulation of these functions as follows:

a. For numbers given as correct to k significant digits the mantissas of their logarithms should have $(k + 1)$ digits; the last digit on the extreme right in this case will represent the apparent error of the logarithm.

b. For approximate angles given as correct to 1′, the mantissas of their logarithmic trigonometric functions should have five places; the last place on the extreme right in this case will represent the apparent error of the logarithmic trigonometric function.

In case the errors are not given in the textbooks, the apparent error of the logarithm is assumed to be ±0.5 of the last place on the extreme right of the mantissa.

While performing Approximate Computation by means of logarithms the pupils should always obtain the relative errors of all the approximate data with which they work, and then by means of the relative error of the final numerical result (which may be numerically equal to the sum of the respective relative errors of the approximate numbers) and the first few significant digits on the extreme right of the final numerical result, they can obtain the apparent error of the final numerical result. In case of logarithmic trigonometric functions this procedure may be found to be too difficult. Still, in case logarithms are used, by means of interpolation and the last digit on the extreme right of the mantissa of the logarithm of the final numerical result, the apparent error of this result can be obtained, as seen from the following example.

Given:

$$a \approx 15.3 \pm 0.05, \quad b \approx 26.4 \pm 0.05, \quad \text{and} \quad B \approx 37^\circ 15' \pm 0'.5.$$

Find A.

We have then

$$\text{sina}\, A \approx \frac{a \,\text{sina}\, B}{b}.$$

Then

$$\begin{array}{r@{\;\approx\;}l}
\text{loga}\, a & 1.1847 \\
\log \text{sina}\, B & 9.7820 - 10 \\
\hline
\log (a \,\text{sina}\, B) & 0.9667 \\
-\text{loga}\, b & 8.5784 - 10 \\
\hline
\log \text{sina}\, A & 9.5451 - 10.
\end{array}$$

Now, since all the logarithms are approximate numbers, it follows that log sina A may have only three digits in its mantissa because two approximate numbers were added and one subtracted, and the apparent error of the final numerical result is 0.00015.

Therefore,

$$\log \text{sina}\, A \approx 9.545 \pm 0.0003.$$

$$\log \text{sina}\, 20^\circ 40' \approx 9.548, \quad \text{and} \log \text{sina}\, 20^\circ 30' \approx 9.544.$$

Then

$$A = 20°30' \pm 3'.$$

In the above computations we used logarithms of a and b with four digits in their mantissas because log sin$a$ 37°15′ had four places in its mantissa, and by such a procedure an optimum approximativeness of the final numerical result is assured.

PART II

THEORY OF NUMERICAL PROCESSES

CHAPTER XIII

INTRODUCTION TO PART II

The Nature of Approximate Computation

The problem of Approximate Computation covers a wide field in its applications to mathematics. It is elementary only so far as the following arithmetical (or numerical) processes are concerned: addition, subtraction, multiplication, division, raising to powers, extraction of roots, and the use of logarithmic and trigonometric functions. Approximate Computation is not a problem in arithmetic only. It is an important phase of applied mathematics, and therefore it does not consist of manipulations with numbers which are at the will and the whim of the computer. It should not and cannot be assumed that the data given in problems that require Approximate Computation are purely numerical. This assumption is fallacious, and therefore if we disregard the fact that the numerical data are measurement data, and are approximate only by virtue of this fact, we may be led astray.

One of the cardinal principles of Approximate Computation is that the errors in and of measurements are not numerical errors only. The errors in and of measurements have a certain property which may be called the accumulative property. The errors that result from measurement cannot be compensated by other errors. A numerical result of Approximate Computation cannot generally be approximated better than the approximate numerical data on which it is based.[1] For example, two measurements, (342.3 + 0.03) yards and (487.5 − 0.03) yards, if added, do not make (820.8 ± 0.00) yards. The statements (820.8 + 0.00) or (820.8 − 0.00) tell us that the number 820.8, as far as the information on hand is concerned, is exact, and there is no approximation at all. In other words, we added two approximate numbers, and we obtained an exact number. This cannot be done. Finite numerical processes with approximations never lead to exact numbers.

This accumulative principle means in other words that: "Errors should never be subtracted from one another. They should always be added." This is the reason why all the errors in Approximate Computation, while we work with them, should be considered in their absolute values. The directions of the errors are very important in the final stage of the process of Approximate Computation.[2]

---

[1] The one exception is to be found in some cases of extraction of roots.

The analysis of the approximativeness of numerical results involves the *combined* use of the apparent error and the relative error. (See page 127). These two types of errors, although their functions differ, are closely related in this analysis, and numerical results should be consistent with their properties. The neglect of this fact leads to ridiculous numerical results; for example, the illustration given on page 156. It must be remembered that the relative error of a product of two (or more) approximate numbers must be greater in its absolute value than any of the relative errors of the factors. If we disregard the accumulative principle, we have the following results, each given as correct to the last significant digit:

$$
\begin{aligned}
(12.52 + 0.004)(6.34 - 0.002) &= 79.377 + 0.0002,\\
(12.52 + 0.004)(6.34 - 0.003) &= 79.4 \quad - 0.03,\\
(12.52 + 0.003)(6.34 - 0.002) &= 79.37 \;\; + 0.001.
\end{aligned}
$$

The criteria for approximativeness, i.e., the apparent error, the relative error, or the significant digits approximativeness should never be mixed beyond the limits required by the numerical processes. The functions of these criteria are different, and unless discretion is exercised, no proper numerical result can be expected. *The decimal place approximativeness, since it conveys no information relative to precision or accuracy, should never be used either in Approximate Computation or in stating values of measurements.*

The rational number system, which is intrinsically limited in the means of representation of the real number system in its totality, is utilized in the process of representation of approximate numbers in this work. The use of approximate numbers pre-dates the theoretical extension of the domain of real numbers to include irrational numbers. The nature of approximate numbers should find its explanation in the nature of irrational numbers. The complexity of the nature of the processes of Approximate Computation should be explained by the fact that an approximate number is generally an approximate representation of an irrational number in a limited rational form[3]. The true nature of the results of Approximate Computation should rest within the nature of this form of representation of irrational numbers. Numerical processes with rational numbers which are tolerably rational only on account of the limitations of the rational number system in common use, but which, since they otherwise represent numbers (or measures expressed in terms of these numbers), possess all the properties of irrational numbers, are more complex than the numerical processes with purely rational numbers. Whatever limitations are imposed on the respective processes, these limitations are logical consequences of the nature of such a number representation.

---

[2] This should not be construed as a departure from the results that were obtained by Laplace and Gauss on the Law of Errors. See secs. IV and V of Chapter XXII for a detailed discussion of this law.

We shall not dwell here upon the theory of irrational numbers. The reader is referred to the works of Weierstrass,[4] Dedekind,[5] and Georg Cantor.[6] The Dedekind and Cantor theories of irrational numbers (which are equivalent; Hobson: *op. cit.*, pp. 40–42) are employed in the development of the theory of Approximate Computation. The only restriction imposed on the processes with approximate numbers which is consistent with the theory of irrational numbers is as follows: *Numerical processes with approximate numbers should take place with approximate numbers belonging to the same class (section) arising in the process of the Dedekind cut.* This restriction, although apparently artificial, is proved upon closer examination of the nature of irra-

---

[3] Approximate numbers which arise from terminating decimals such as, for example,

$$\frac{1}{8} = 0.125, \quad \frac{1}{32} = 0.03125, \text{ etc.},$$

when the number of the digits in the terminating decimal is limited, for example,

$$\frac{1}{32} = 0.03 + 0.002,$$

fall in this class also.

[4] His lectures on Analytical Functions. An exposition of this theory of Weierstrass is also found in: Biermann, O.: Theorie der analytischen Funktionen. Leipzig, 1887, p. 19.

[5] Dedekind, R.: Stetigkeit und irrationale Zahlen. An English translation by Beman, W.: Essays on Number. Open Court Publishing Co., Chicago, Ill., 1901.

[6] Mathematische Annalen, Vol. V. (1872), p. 123.

Mathematische Annalen, Vol. XXI (1883), p. 545. In the last paper Cantor discusses the three theories.

An account of the Dedekind and Cantor theories of irrational numbers is found in:

Hobson, E. W.: The Theory of Functions of a Real Variable. Vol. 1, Chap. I., Cambridge University Press, 1927.

Pierpont, J.: The Theory of Functions of Real Variables. Vol. 1, Chap. II (secs. 54–80 and secs. 128–131). Ginn & Co., Boston, 1905.

tional numbers to be entirely justifiable, and there is nothing arbitrary in it.[7] We shall return to this restriction presently.

If we denote an approximate number by a, the rational representation of the approximate number is then either

$$a = A_1 - e,$$

or

$$a = A_2 + e',$$

where A_1 and A_2 are rational numbers, and e and e' are the corresponding *apparent errors* in and of the measurements, and

$$|e| + |e'| = A_1 - A_2.$$

In the irrational number system, according to the Dedekind mode of representation, we would have the irrational number A (the "cut"), and A_2 would belong to the lower aggregate, and A_1 would belong to the upper aggregate, that is,

$$0 < A_2 < A < A_1.$$

Accordingly, we introduce the *absolute error* of a, α, which corresponds to A, and we have

$$|a| \leqslant |A| + |\alpha|.$$

Whatever numerical processes take place, the rational representation of the irrational number should preserve its nature, so that the presence of the Dedekind cut and of the two distinct aggregates may be retained. This is effected if and only if the approximate numbers have rational parts (which are otherwise known as *the limits or bounds of the true value*) of the same nature. Operations with rational parts (and consequently with apparent errors) belonging to aggregates of opposite nature will necessarily upset the entire structure of the representation of irrational numbers. Consequently, the final numerical result of an operation thus performed will have no sense.

At any given stage of approximation the absolute error of an approximate number a approaches a limit which is known as *the upper limit of the absolute error.* Thus

$$|\alpha + \epsilon| \sim |\alpha_1|, \quad (\epsilon \rightarrow 0).$$

---

[7] Gauss sensed the necessity of such a restriction. He suggested that the errors of approximate numbers be affected by the same sign. See his Theoria Motus Corporum Coelestium, Hamburg, 1807, sec. 31.

The infinitesimals used here in the process of analysis are to be distinguished from the infinitesimals used in the Function Theory.

> It should be remarked that, in assuming that every section of the aggregate of real numbers defines a single real number, it has been implicitly assumed that *if a, b are any two positive real numbers, such that $a < b$, then a positive integer n can be found such that $na > b$.* This is the arithmetical analogue of the so-called principle of Archimedes. . . .
>
> A consequence of the fact that the aggregate of real numbers forms an Archimedean system is that so-called infinitesimal numbers do not exist within the aggregate. Every positive number ϵ, being such that an integer n can be found such that $n\epsilon > 1$, is a finite number, in the sense in which finite numbers were distinguished from infinitesimals in the older forms of the Infinitesimal Calculus. *In Arithmetical Analysis the conception of the actually infinitesimal has no place.* When the expression "infinitesimal" is used at all, it is to describe the process by which a variable to which the numbers of a sequence converging to zero are successively ascribed, as values, approaches the limit zero; thus an infinitesimal is a variable in a state of flux, never a number. Such a form of expression, appealing as it does to a mode of thinking which is essentially non-arithmetical, is better avoided. . . .[8]

The concept of the upper limit of the absolute error is in conformity with Cantor's definition of an irrational number, and the irrational number A is thus rationally represented as a convergent sequence whose terms are monotonically decreasing.[9] The concept of the upper limit of the absolute error is important from the ordinal point of view. It can be shown that the processes of addition and subtraction of approximate numbers should take place with approximate numbers whose absolute errors are of the same order. The orders of the absolute errors are directly tied up with the positional location of their respective apparent errors, and this leads directly to the concept of rounding off approximate numbers. The equality of the orders of the absolute errors is determined from the relationship

$$\frac{|\alpha_1|}{|\beta_1|} \underset{\sim}{=} 10^{-r}, \quad (r = 0, 1, 2, 3, \ldots),$$

when $r = 0$.[10]

Another criterion for the approximativeness is the relative error, which is defined as

$$\varphi_a = \frac{\alpha}{A},$$

and the upper limit of the relative error. The upper limit of the relative error, together with the first significant digit on the extreme left of the true value portion of an approximate number and the significant digit of the apparent error of the approximate number, offers a perfect criterion for the determination of the approximativeness of a number, and this is known as the combined apparent and relative error test.[11] This test is important in the entire theory of Approximate Computation.

---

[8] Hobson: *op. cit.*, pp. 42–43.

[9] "Monotonic" describes a variation in one direction only, that is, a function does not oscillate in the interval for which it is described.

[10] The sign $\underset{\sim}{=}$ reads "belongs to the order."

The ordinal property of the relative errors of approximate numbers is very important, and this property should be employed throughout the analysis of the processes of multiplication, division, raising to a power, and extraction of roots. By means of the test described above and the ordinal property of the relative error, the analysis of the four numerical processes mentioned is effected.

The absolute and relative errors are exceedingly important in the analysis of the properties of logarithmic and trigonometric functions. These functions, being transcendental, are represented by numerical values (with a very few exceptions) that are irrational, and hence approximate. There is, however, a distinct relationship between the approximativeness of the argument and the approximativeness of its corresponding function, and this relationship is fully developed in this work.

The method of analysis pursued in this work has as its aim the development of criteria for approximativeness and rules for Approximate Computation which are ultimately stated in terms of significant digits. This last criterion implies the assumption that all the digits of a result of computations with approximate numbers located in places free from errors are significant regardless of the magnitude of the first digit (not zero) on the left of the apparent error of this result. Rounding off the result is a second step in Approximate Computation, and this process is not incorporated in the statements of rules in terms of significant digits.

In order that, after the numerical processes have taken place, a computer should obtain the proper results, that is, results in the form he needs, he should be aware of the difference between:

1. *the exact* and *the approximate.*
2. *correctness* and *approximativeness.*
3. *accuracy* and *precision.*

Great care should be exercised in the so-called "rounding off" of digits. The method of rounding off digits as used in the case of pure numbers has no place in the processes of Approximate Computation. Approximations represent measurements, and the idea of the actual measurement should always accompany the reasoning in the process of rounding off digits. An approximate number is usually represented by means of one of the two limits of the true value coupled with its corresponding apparent error (if such is given, otherwise the apparent error is assumed to be ±0.5 of a unit of the last significant digit on the extreme right of the number representing the true value portion of the approximate number). One should question the justification for rounding off (unless this is required for some definite reason) the approximate number 2,106.5 + 0.06 to 2,106. + 0.6. This should be done when the analysis of the errors calls for an error in the fifth place. Otherwise, the proper rounding off is 2,106.6 − 0.04. Overzealousness in rounding off digits may lead to unnecessary loss in accuracy and precision. On the other hand, looseness in thinking and careless retention of unnecessary digits may bring no less disastrous results.

---

[11] See Chap. XIV, page 127.

The theory of Approximate Computation, as developed in this work, is treated purely mathematically, but without neglecting the fact that the approximate data under consideration are measurement data and are approximate by virtue of this fact. Much use has been made of symbolism, and a list of the symbols used in this work is found in Appendix 3.

CHAPTER XIV

APPROXIMATIONS

1. EXACT AND APPROXIMATE QUANTITIES

IN PRACTICAL mathematics two classes of quantities are to be distinguished, one the antithesis of the other; namely, *exact* quantities and *approximate* quantities. Both classes may be represented by numbers: exact numbers and approximate numbers respectively.

There is one exception, however. Certain exact numbers do not have corresponding approximate numbers, and many (if not most) of the approximate numbers do not have corresponding exact numbers, unless such exact numbers are thought of as "ideals" in the Platonic sense.

For example, the result of the actual count of the pupils in a class is represented by an exact number. There is no sense in an approximation. The approximate number in such a case, for instance, (42.5 ± 0.05) pupils, is absurd. There cannot be seven and one-half live fishes in a pond. Should these two numbers represent averages, however, there would be a definite significance in an approximate (and fractional) representation of "indivisible units." The numbers π, $\sqrt{2}$, e, and the like, are irrational numbers. They cannot be represented *exactly* in terms of the rational number system. They can be expressed approximately only, and to only a certain degree of approximativeness. Still π, $\sqrt{2}$, log 3, and the like, can be thought of as *ideals* without considering their respective approximate numerical values. Likewise, many fractions cannot be expressed as terminating decimals; non-terminating decimals representing such fractions are merely approximations (when such a decimal is written to a certain number of digits at the discretion of the computer). *Hereafter only decimal fractions will be considered and used in this book.*

Approximate quantities, unless they are represented by numerical values of certain types of mathematical functions (such as irrational and transcendental) which are obtained from tables (or by any other means) are results of measurements. The idea of *measurement* is used here in its broadest sense. Any result of an observation which is stated numerically is a statement of measurement. It may be a school grade, a temperature, a length, a percentage, an average, the size of a deviation, or a ratio. It may be a statement of a direct measurement. It may be a statement of an indirect measurement.

II. THE APPROXIMATIVENESS

Let $y = f(x)$ be defined and be continuous within a certain interval (x_0, x_n), end points included, all the values of x_k $(k = 0, 1, 2, \ldots)$ being real. Then at the points x_k the function $y = f(x)$ has the values $y_k = f(x_k)$.

If the values of $\overline{y_k}$ $(k = 0, 1, 2, \ldots)$ that are obtained from the equation $y = f(x)$ by substituting in it the values of x_k $(k = 0, 1, 2, \ldots)$ successively (with no regard to the order of the substitution of x_k) do not deviate from the actual values of y_k, then the values of $\overline{y_k}$ are known as *the exact values.* In other words,

$$\overline{y_k} \equiv y_k \quad (k = 0, 1, 2, \ldots)$$

denotes the exact value.

If the values of $\overline{y_k}$ deviate from y_k, regardless of the direction of the deviation, the values of $\overline{y_k}$ are said to *approximate* the values of y_k $(k = 0, 1, 2, \ldots)$, that is,

$$\overline{y_k} \sim y_k \quad (k = 0, 1, 2, \cdots)$$

denotes *the approximate value.*

The closer the values of $\overline{y_k}$ are to the values of y_k the closer is the approximation, or the greater is the approximativeness. The closeness of the approximativeness depends on how closely

$$|y_k - \overline{y_k}| \quad (k = 0, 1, 2, \cdots)$$

approaches zero. Hence, if

$$|y - \bar{y}| < |\alpha_i| \quad (\alpha_i \to 0; \quad i = 0, 1, 2, \cdots)$$

$\bar{y}$ is said to be the approximation of y with a certain degree of approximativeness. α_i in this case is the *excess* (positive or negative) of the approximation $\bar{y}$. The excess is commonly known as *the error.*

III. FORMULAS

Practical computation involves the application of formulas to numerical data. The approximativeness of numerical results of performed computations is always based on two criteria: (a) the approximativeness of a formula, and (b) the approximativeness of the numerical data.

The approximativeness of a purely mathematical formula may sound paradoxical from the start, but a closer analysis reveals very important properties of these formulas.

A. EXACT MATHEMATICAL FORMULAS

A purely mathematical formula (algebraic, geometric, trigonometric, etc.) is a statement of a certain relationship in a purely mathematical sense. As far as the statement of the relationship is concerned such a formula may be "ideal," that is, exact.

The expression for the roots of the general quadratic equation $ax^2 + bx + c = 0$:

$$x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a},$$

the following formulas:

$$\sin 2x = 2 \sin x \cos x,$$

$$V = \frac{4}{3} \pi r^3,$$

and the like, are examples of mathematical formulas. These formulas are exact in the mathematical sense.

Exact mathematical formulas may involve certain operations or certain quantities that will necessarily render the numerical results of performed computations approximate, *even though all the numerical data are exact.* If such formulas in final form, after numerical substitutions have been made, contain:

1. ratios represented by irreducible fractions whose denominators contain as factors powers of integers other than 2 and 5 (and very often these are not excepted);
2. irrational numbers such as π, e, logarithmic functions, and trigonometric functions (except those in which the values of the logarithmic and trigonometric functions are expressed by exact numbers in the rational number system, provided the arguments of these functions are also exact);
3. powers of certain expressions, such as the compound interest formula;
4. surds, unless the number under the radical sign is a perfect power of an integral multiple of the index of the surd;

then the numerical results of performed computations *are invariably approximate.*

If the numerical data are approximate, then the application of an exact mathematical formula to such data never yields exact results after the computations are performed.

B. APPROXIMATE MATHEMATICAL FORMULAS

A mathematical formula may be approximate in the mathematical sense. Some of such formulas are found in elementary mathematics. For example,

1. The interpolation formula (the first terms only):

$$f(x) = f(a_0) + \sum_{1}^{n} \frac{x^{(k)}}{k!} \Delta^k f(a_0) + R_n,$$

where

$$x^{(k)} = x(x-1)(x-2) \cdots (x-k+1),$$
$$\Delta f(a_0) = f(a_0 + 1) - f(a_0),$$

and the remainder R_n expresses the error of the computation to the nth difference.

2. Simpson's rule which is written in general form as

$$\int_a^b f(x)dx \sim \frac{h}{3}\Big\{f(a) + 4f(a+h) + 2f(a+2h) + 4f(a+3h) + \cdots$$
$$\cdots + 4f[a+(n-1)h] + f(b)\Big\}.$$

Approximate mathematical formulas yield numerical results which are only approximate. In the case of formulas of the type illustrated by the two formulas above the greater the number of terms used in the formula in the process of computation, the greater is the approximativeness attained. In the case of Simpson's rule the number of units depends on the size of the unit $h = \frac{b-a}{n}$.

C. SCIENTIFIC FORMULAS

Formulas of the third type are neither exact nor approximate in the sense of a purely mathematical formula. These are scientific formulas, and they are developed as results of experimentation (including observation); for example, the statement of Boyle's law

$$pv = k,$$

where p denotes pressure,
v denotes volume,
and k is a certain constant determined approximately by experiment.

The above relationship gives a reasonable approximation to the true state, provided the gas is not compressed too much. If v is decreased, and p is increased beyond a certain point, the above relationship ceases to be even approximate. A better approximation is given by van der Waal's law

$$\left(p + \frac{a}{v^2}\right)\left(v - b\right) = c,$$

where a, b, and c are certain constants which are determined approximately by experiment. Neither of these two laws expresses, however, the exactly true relationship. They are merely "tolerable" expressions that state the approximate relationship.[1] We can contemplate from a mathematical point of view an ideal formula that would express the relation between p and v for all possible cases; but physical sciences have not yet reached the stage where relations of that type can be formulated.

Consider the problem of the simple pendulum. The formula for the period of oscillations is commonly given as

$$P = 2\pi\sqrt{\frac{l}{g}},$$

where l is the length of the pendulum,
and g is the gravitational constant whose value varies with the latitude on the earth's surface.

The fact is that this formula is valid for infinitesimal oscillations only, on the basis of which the formula is derived. For finite (non-infinitesimal) oscillations the approximation is

$$P = 2\pi\sqrt{\frac{l}{g}}\left\{1 + \left(\frac{1}{2}\right)^2 k^2 + \left(\frac{1.3}{2.4}\right)^2 k^4 + \dots\dots\dots \right.$$
$$\left. \dots\dots\dots + \left[\frac{1.3.5.\dots\dots\dots(2n-1)}{2.4.6.\dots\dots\dots 2n}\right]^2 k^{2n} + \dots\dots\dots\right\},$$

where $k^2 = \sin^2 \tfrac{1}{2}\theta,$

and θ is the angle of oscillations.[2]

The application of formulas of the above type to numerical data, whether approximate or exact, must invariably yield numerical results that are approximate only.

---

[1] This example is borrowed from G. H. Hardy: A Course in Pure Mathematics. Cambridge University Press, 1928, p. 39.

[2] Macmillan, W. D.: Theoretical Mechanics; Statics and Dynamics of a Particle. McGraw-Hill Book Company, Inc., New York, 1927, pp. 311–312

IV. Criteria for Determination of Approximativeness

Formulas of the above three types are common in practical work in mathematics. In most of the cases in practical mathematics the results of the application of such formulas to numerical data are approximate. It is certainly imperative then that the application of a formula should never be merely a mechanical process. It is probably just as important as the correct performance of the computation that the nature of the results of the computer should be determined by him.

The analysis of a formula would not indicate the actual numerical limits of approximativeness by which the computer could be guided. The determination of the limits of approximativeness of the numerical results of the application of a formula to numerical data involves besides the knowledge of the nature of the formula specific knowledge of the approximativeness of the data.

V. Numerical Data

Numerical data may be either exact or approximate. Data relative to counting (unless allowance is made for slips in the counting process) are always exact and are usually represented by exact numbers. "Number (of discrete individuals) is absolute. It is the result of counting, and counting is an absolute operation. If two men count the number of people in this room and reach different results, one of them must be wrong."[3] The number of pupils in a school, the amount of money spent, the number of days in a given period of time, are data of this type. Data of this type are not necessarily represented by integers only; for example, two and a half dollars.

Data secured from tables of numerical constants giving the values of irrational or transcendental functions are with a very few exceptions approximate.

Data relative to measurements are always *approximate*, for three reasons:

1. There is no measuring instrument, however precise, that can measure exactly. Each measuring instrument has its individual defects which cause errors of measurement, and these errors are not generally constant. They change with the temperature of the day, with the barometric pressure, with the humidity of the air, and so

---

[3] Eddington, Sir A. S.: The Nature of the Physical World. Cambridge University Press, 1932, p. 23.

forth. These errors must be taken into consideration whenever measurements are made, and the results of the measurements must be *corrected* to these errors.

2. The units of measure themselves are only approximately known. First of all they must be expressed in terms of some other units of measure, and these units must be measurable in terms of some other units, and so on—an endless chain. For example, the standard meter as given by Michelson (1907) in terms of the wave length λ of the red light in the spectrum of cadmium at 760 *mm.* (millimeters) barometric pressure in absolutely dry air at 15° centigrade:

$$1\ m. = 1{,}553{,}163.6\lambda, \quad \text{and} \quad \lambda = 0.64384722\mu,$$

where

$$\mu = 0.001\ mm.$$

Secondly, certain units of measure, although expressed in terms of some other units of measure, vary from time to time. For example, considering two successive passages of the sun over the meridian, the day is never 24 hours long. The year is not exactly 365 days long.

3. The errors that arise in the process of measurement are unavoidable errors due to the individual characteristics of those who perform the measurements. Two persons applying the same yardstick to the same object will never obtain exactly the same length. Certain observers have their individual characteristics. Some of them set their instruments with less care. Some of them make slight errors in reading, and so on.

The above considerations lead to the conclusion that *any numerical data of measurement are at the most as nearly perfect* in their approximation of the exact measure as the precision of the instrument of measurement allows. This limitation is of prime importance in Practical Mathematics, and in the interpretation of the results of the application of formulas to numerical data.

Approximate data supplied by tables of logarithmic and trigonometric functions as well as other tables (squares, cubes, square roots, cube roots, etc.) are subject to the limitations of the tables themselves. Otherwise, such data are governed in their approximativeness only by the means by which they are computed.

VI. APPROXIMATE NUMERICAL DATA

Exact values can be represented by all rational and irrational numbers. Whether a number is exact or approximate depends on the interpretation of the data.

For example, the length of a diagonal of a cube whose edge is exactly 1 ft. is $\sqrt{3}$ ft., and $\sqrt{3}$ ft. in this case is the exact length of the diagonal of the cube, but it is an "ideal" as stated in section I above. If, however, it is required to measure this length with a ruled tape, no such length can be obtained, and it is necessary to approximate.

If 5 lbs. of sugar cost 29¢, the price of 1 lb. is 5.8¢. If 1 lb. is purchased there are no coins which would in any combination make 5.8¢. In this case 5.8¢ is exact, but in actual experience an approximation (6¢) is used.

VII. THE CRITERION FOR APPROXIMATIONS

The criterion for approximations

$$\overline{y_k} \sim y_k \qquad (k = 0, 1, 2, \cdots)$$

$$|y - \overline{y}| < \alpha_i \qquad (\alpha_i \to 0; \quad i = 0, 1, 2, \cdots)$$

given in section II, can be translated into numerical form.

The decimal number system that is used for number representation is a *positional* system. Each number digit a_k ($a_k = 0, 1, 2, \ldots, 9$; $k = 1, 2, 3, \ldots$) occupies a definite position in the number

$$10^p(10^n a_1 + 10^{n-1} a_2 + 10^{n-2} a_3 + \cdots + 10 a_{n-1} + a_n),$$

where n is an integer > 0,
and p is an integer $\gtreqless 0$.

By virtue of its position the digit a_k is *significant* in the sense that it denotes a certain magnitude. It is significant also because it occupies a place of significance which is directly connected with the fact that it relates some information regarding that magnitude. The transposition of a certain digit to the left of its original position raises its importance, for it denotes then a magnitude ten times as large as before, whereas the transposition of the digit to the right of its original position lowers its importance, for then it denotes a magnitude ten times smaller. By virtue of this property, zero (0) is significant only when it is located between two other digits that occupy places of significance and are significant. Several zeros may be thus significant if all of them are located between two significant digits. Zero is also significant when it is on the extreme right of a number, and there is definite indication that its presence there is justified by virtue of the performed measurement or computation. Thus, the last zero on the extreme right of log 2 = (0.3010 + 0.00003) is significant.

The approximativeness of a number according to the above criterion must, therefore, be determined by consideration of certain facts that occur on the right side of a multi-digited number. Comparison between two numbers, one representing an exact value (or *the true value*), and the other its approximation, must reveal that up to a certain position the digits are identical, and to the right of that position the digits are different. For example,

$$10^p(10^n a_1 + 10^{n-1}a_2 + \cdots + 10^{n-k+1}a_k + 10^{n-k}a_{k+1} + \cdots + 10a_n + a_{n+1})$$

and

$$10^p(10^n a_1 + 10^{n-1}a_2 + \cdots + 10^{n-k+1}a_k + 10^{n-k}b_1 + 10^{n-k-1}b_2 + \cdots)$$

The two numbers, the first being exact and the second approximate, agree up to and including the kth place from the left. For example, take the number

$$\pi = 3.14159\ 26\underline{535} \cdots,$$

and its approximation

$$\pi = 3.14159\ 26\underline{72} \cdots;$$

the digits between which there is disagreement are underlined. These numbers agree up to and including the place of ten-millionths, or it can be said that they agree in eight places. In other words, the approximation is correct to eight significant digits. It should be noted that the approximation exceeds the true value of π.

The above illustration shows that an approximation can be carried out in the following manner:

1. A number can approximate the exact value to a stated unit.
2. A number can approximate the exact value to a stated number of significant digits.

The excess (positive) or the defect (negative) of the approximation over or under the exact value is known as the *error* of the approximation, and the value of this excess or defect is necessarily a fraction of a unit in the place of significance of the last digit on the extreme right to which the exact number and the approximate number agree. It follows then that:

3. A number can approximate the exact value within a given error or excess (positive or negative).

VIII. THE DECIMAL PLACE APPROXIMATIVENESS

The approximation of exact values by numbers correct to a certain number of decimal places is common to physical measurements. Instruments of measurement, and especially instruments of precision, are usually known as being able to measure to a decimal fraction of a given unit of measure. For instance, a vernier may be graded so that it gives readings to one tenth of a unit of measure. This means that the readings of the tenths on the vernier are correct insofar as the information which is imparted by such readings indicates that the obtained measure varies from the exact measure by some fraction of the tenths. For example, the true length of a diagonal of a square whose side is one foot is

$$(1.4142 \cdot \cdot \cdot \sim \sqrt{2}) \text{ ft.}$$

The reading on a yardstick with a vernier that reads to tenths of a foot is 1.4 ft., and it is correct to tenths, the error being in the hundredths, i.e., 0.01 + (the sign + indicates that there is an error in the thousandths, etc.).

The use of the method of representing approximate numbers correct to a stated number of decimal places brings the result that all numbers which are given to the same number of decimal places have errors whose numerical values are all of the same order, i.e., in hundredths, thousandths, etc.

IX. THE SIGNIFICANT DIGITS APPROXIMATIVENESS

Approximation of values by numbers that are correct to a stated number of significant digits, however, dispenses with the precision element of instruments of measure, to some extent.

The underlying principle of this method of representation of approximate numbers is based on the uniformity of the orders of certain ratios.[4] These ratios are ratios of the errors to the true values of their corresponding approximate numbers, and in the case of approximate numbers correct to the same number of significant digits these ratios are "tolerably" of the same order. The assignment of these ratios to their respective orders is accomplished by means of inequalities. For example,

$$\log 2 = 0.3010 + 0.00003,$$
$$\log 9 = 0.9542 + 0.00004.$$

Then these ratios are

$$\frac{3}{30{,}103} \text{ and } \frac{4}{95{,}424}.$$

---

[4] See section XVI below.

and the assignment to their respective orders is indicated as follows:

$$0.00001 < \frac{3}{30,103} < 0.0001,$$

$$0.00001 < \frac{4}{95,424} < 0.0001.$$

We may have two approximations

$$\log 3 = 0.4771 + 0.00002,$$

$$\log 6 = 0.6989 + 0.00007,$$

which are both correct to the same number of significant digits, but their assignment to the respective orders would show that the ratio corresponding to log 6 is greater than the ratio corresponding to log 3. This is the reason why the above statement was qualified by the use of the word "tolerably." We have then

$$0.00001 < \frac{2}{47,712} < 0.0001,$$

$$0.0001 < \frac{7}{69,897} < 0.001.$$

These ratios are very important in the process of Approximate Computation.

X. THE APPROXIMATIVENESS OF NUMERICAL RESULTS

The representation of approximations within a certain given error, where no reference is made to the number of significant digits or to the number of decimal places, combines the two methods described above. Apart from these two methods, the use of errors as a criterion of approximativeness is important only from the point of view of evaluating the approximativeness of the results of application of formulas to numerical data. In some final stages of Approximate Computation, however, the representation of approximate numbers should revert to the indication of the number of significant digits.

XI. THE APPARENT ERROR

Consider the exact number A

$$A = 10^p(10^n x_1 + 10^{n-1}x_2 + \cdot\cdot\cdot + 10^{n-k+1}x_k + 10^{n-k}x_{k+1} + \cdot\cdot\cdot)$$

where

n is an integer > 0, k is an integer > 0,

p is an integer $\gtreqless 0$, $p \gtreqless n$, and

$$x_1 \neq 0, \quad \text{and} \quad x_i = 0, 1, 2, \cdots, 9; \quad i = 2, 3, 4, \cdots$$

In actual practice, when an exact number cannot be represented in the rational number system, or when a measurement is performed, there can be two cases: either

$$a = 10^p(10^n x_1 + 10^{n-1}x_2 + \cdots + 10^{n-k+1}v_1 + 10^{n-k}v_2 + \cdots),$$

or

$$a' = 10^p(10^n x_1 + 10^{n-1}x_2 + \cdots + 10^{n-k+1}w_1 + 10^{n-k}w_2 + \cdots),$$
$$v_1 < x_k < w_1.$$

(We shall assume the values v_1, x_k, and w_1 as having been rounded off or forced when necessary.)

In most of the cases the exact value A is such that it has to be considered as ideal (see section 1). Since $a \sim A$, and A is located between the two values a' and a, these two values are the approximations of A. For example, such is the expression for π as given by Vieta (1540–1603):

$$3.14159\,26535 \cdots < \pi < 3.14159\,26537 \cdots$$

Either of the numbers

$$10^p(10^n x_1 + 10^{n-1}x_2 + \cdots + 10^{n-k+1}v_1),$$
$$10^p(10^n x_1 + 10^{n-1}x_2 + \cdots + 10^{n-k+1}w_1)$$

will be hereafter known as *the true value*. The remaining portions of a and a', namely

$$10^p(10^{n-k}v_2 + \cdots) \quad \text{and} \quad 10^p(10^{n-k}w_2 + \cdots)$$

are known as the excesses (or errors).

Consider the two equivalent forms

$$10^p(10^n x_1 + 10^{n-1}x_2 + \cdots + 10^{n-k+1}v_1 + 10^{n-k}v_2)$$

and

$$10^p[10^n x_1 + 10^{n-1}x_2 + \cdots + 10^{n-k+1}(v_1 + 1) - 10^{n-k}(10 - v_2)],$$

where v_2 is rounded off. These two numbers represent the same measure with the only difference that in one case the excess is positive, and in the other case the excess is negative. We shall agree that whenever it is convenient the excess (or error) should be rounded off[5] so that $|v_2| \leqslant 5$.

---

[5] See Appendix 1 on "Rounding Off Digits."

This allows the introduction of the notation of the *limit of the error.*

1. If an approximate number is given as correct to k significant digits, then its error is *at the most* equal to ± 0.5 of a unit in the kth place, counting from the left to the right.

2. If an approximate number is given as correct to a $10^m (m \gtreqless 0)$ unit, then its error is *at the most* equal to $\pm 0.5 \cdot 10^m$.

Let us denote the exact value by A. Denote by

$$A_2 = 10^p(10^n x_1 + 10^{n-1} x_2 + \cdot \cdot \cdot + 10^{n-k+1} v_1),$$
$$A_1 = 10^p[10^n x_1 + 10^{n-1} x_2 + \cdot \cdot \cdot + 10^{n-k+1}(v_1 + 1)].$$

Then

$$A_2 < A < A_1.$$

A_1 will be hereafter known as *the upper limit* (or *upper bound*) *of the true value.* A_2 will be hereafter known as *the lower limit* (or *lower bound*) *of the true value.* We shall denote by a the approximate number which corresponds to A, A_1, and A_2.

The differences $a - A_2$ and (or) $A_1 - a$ with their respective signs denoting their directions are known as *the apparent errors* of the approximate number a.

XII. THE EXACT VALUE OF AN APPROXIMATE NUMBER AND THE ABSOLUTE ERROR

The difference $|A - a|$ is known as *the absolute error* of the approximate number a. The absolute value is taken here for $A - a$ because there might be two cases, namely

$$A < a \quad \text{or} \quad A > a.$$

We shall denote

$$|A - a| = \alpha.$$

Since

$$a \sim A,$$

and the closer the approximation,

$$\alpha = |A - a| < |\epsilon| \quad (\epsilon \rightarrow 0).$$

In practical applications the absolute error is always an ideal, and the numerical processes of Approximate Computation involve apparent errors only. We shall make use of the absolute errors for generalization purposes, however.

Considering the difference $|A - a| = \alpha$ at a given stage of approximation we note that the absolute error α approaches an upper limit α_1, and it approaches it asymptotically. For example, consider an approximation of π to three significant digits, 3.14. For this approximation we have a series of absolute errors: 0.001, 0.0015, 0.00159, 0.001592, 0.0015926, etc. We have then

$$\alpha \sim \alpha_1,$$

or

$$\alpha + \eta = \alpha_1, \quad (\eta \rightarrow 0).$$

Moreover, since $a \sim A$, by virtue of the above it may be said that a approximates the value of A with an absolute error $\alpha < \alpha_1$. We have then the inequalities:

$$A - \alpha < a < A + \alpha,$$
$$A - \alpha_1 < a < A + \alpha_1,$$

and

$$a - \alpha_1 < A < a + \alpha_1.$$

Since

$$A_2 < A < A_1$$

we have

$$A_2 < a < A_1.$$

Apparent errors of final numerical results should always be changed to the upper limits of their corresponding absolute errors. In other words, if an apparent error is obtained such that it has several significant digits, its first significant digits on the extreme left should be forced regardless of the magnitude of the significant digit that is on the right of it.

XIII. THE RELATIVE ERROR

The ratios $\frac{a - A_2}{A_2}$ and $\frac{A_1 - a}{A_1}$ are known as the respective *apparent relative errors* or the relative errors of the approximate number a. It shall be agreed to compute the relative error of a for the apparent error when this apparent error is ± 0.5 of a unit of the last place of significance on the extreme right of a (unless otherwise specified).

For analysis purposes we shall use the ratio

$$\wp_a = \frac{\alpha}{A}$$

which will be hereafter called the relative error of a.

By virtue of the fact that

$$\alpha \sim \alpha_1$$

and

$$\alpha + \epsilon = \alpha_1, \ (\epsilon \rightarrow 0),$$

we have

$$\frac{\alpha}{A} < \frac{\alpha}{A_2} < \frac{\alpha_1}{A_2}.$$

We also have

$$\alpha > \wp_a A_2$$

In practical applications the so-called *relative error test* is based on the interpretation of the *upper limit* of the relative error. The upper limit of the relative error is derived as follows. Consider

$$A = 10^p(10^n x_1 + 10^{n-1} x_2 + \cdot \cdot \cdot + 10^{n-k+1} x_k + 10^{n-k} x_{k+1} + \cdot \cdot \cdot)$$
$$a = 10^p(10^n x_1 + 10^{n-1} x_2 + \cdot \cdot \cdot + 10^{n-k+1} x_k + \alpha),$$

where

k is an integer > 0, n is an integer > 0,

p is an integer $\gtreqless 0$, $p \gtreqless n$,

$x_1 \neq 0$, $x_i = 0, 1, 2, \cdot \cdot \cdot , 9$; $i = 2, 3, 4, \cdot \cdot \cdot ,$

and

$$\alpha \leqslant 0.5 \cdot 10^{p+n-k+1}.$$

Then

$$\wp_a < \frac{0.5 \cdot 10^{p+n-k+1}}{A}.$$

No attention is paid to the direction of the relative error because the direction of the relative error is the same as the direction of its corresponding absolute error. We have then

$$\wp_a < \frac{0.5 \cdot 10^{p+n-k+1}}{10^{p+n} x_1} < \frac{1}{10^{k-1} x_1},$$

and the upper limit of the relative error $\wp_a$ is defined as

$$\wp'_a < \frac{1}{10^{k-1}}.$$

We can state then: If an approximate number whose first significant digit on the extreme left is x_1 is correct to k significant digits, then its relative error is

$$\wp_a < \frac{1}{10^{k-1} x_1},$$

and the upper limit of its relative error is

$$\wp'_a < \frac{1}{10^{k-1}}.$$

Conversely, if the relative error of an approximate number, whose first significant digit on the extreme left is x_1, is

$$\wp_a < \frac{1}{10^{k-1}x_1}$$

then the approximation is *at the most* correct to k significant digits.

If the upper limit of the relative error of an approximate number is

$$\wp'_a < \frac{1}{10^{k-1}}$$

then the approximation is correct to k or $(k - 1)$ significant digits. This depends entirely on the magnitude of the first significant digit on the extreme left of the approximate number and on the magnitude of the first digit (significant) on the left of the apparent error of the approximate number.

The relation between the first significant digit on the extreme left of the approximate number, the significant digit of the apparent error (or absolute error), and the upper limit of the relative error of the approximate number is given in Table 4, to which constant reference will be made hereafter. Results secured by means of this table are entirely reliable. The computer should make it a practice, however, to check his results. The process of checking the results will enhance the importance and the significance of the interpretation of the approximativeness of numerical results obtained after for-

TABLE 4

The upper limit of the relative error is located between	The number of significant digits of the approximate number	
	The significant digit of the apparent error is $\leqslant x_1$	The significant digit of the apparent error is $>x_1$
0.5 and 0.1	–	1
0.1 and 0.01	1	2
0.01 and 0.001	2	3
0.001 and 0.0001	3	4
0.0001 and 0.00001	4	5
....	.	.
....	.	.
10^{-n} and $10^{-(n+1)}$	n	$n + 1$

mulas are applied to numerical data (whether approximate or exact), and after all the stated numerical operations are performed with the data.

The *relative error test* is, therefore, very important in the processes of Approximate Computation, when the approximativeness of the final numerical results must be evaluated.

Examples

1. The relative error of π correct to 8 significant digits is

$$<0.00000003.$$

2. π is computed with a relative error of 0.0005. The first significant digit of π is 3. The first significant digit of the apparent error is 2. The upper limit of the relative error is 0.001. Hence the approximation of π is safely correct to 3 significant digits.

If a number is computed to n digits, and after the evaluation of its approximativeness it is found that there are only k significant digits $(k < n)$, then $(n - k - 1)$ digits, counting from the right, should be *dropped* (or *suppressed*). The $(k + 1)$st digit, counting from the left, if necessary should be rounded off first, and this $(k + 1)$st digit constitutes the apparent error of the approximate number, or if the apparent error is computed separately, this $(k + 1)$st digit should be added to the computed apparent error.[6]

If the relative error, or its upper limit, is multiplied by 100, then the relative error becomes a percentage error.

XIV. PRECISION AND ACCURACY

There is a significant difference between the function of the absolute (or apparent) error and the function of the relative error. Both errors state the fact that a number (or a measure) is approximate, but they state this fact in different ways. The absolute (or apparent) error tells about *the precision* of a measurement or of a computation. The relative error tells about *the accuracy* of a measurement or of a computation.

Precision and accuracy should never be confused. As criteria for a study of approximativeness these two ideas are widely different, and at times mutually exclusive, and in many cases where precision benefits, accuracy is sacrificed, and vice versa. The following example will illustrate this point: If an object is measured to one onemillionth of an inch, and a measure correct to 3 significant digits is obtained, for example,

---

[6] See Appendix 1 on "Rounding Off Digits."

$$(0.000355 \pm 0.0000005) \text{ in.},$$

the apparent error of the measurement is very small in comparison with the apparent error of a measurement of a distance correct to one foot when a measure is obtained correct to 4 significant digits, for example,

$$(4{,}226. \pm 0.5) \text{ ft.}$$

The first measure is decidedly more precise. But the second measure is more accurate, for the relative errors are respectively 0.0014 and 0.00012. In per cents these errors are 0.14% and 0.012% respectively. Obviously, the second measurement, although its unit is a foot, is the more accurate one.

XV. ABSOLUTE (APPARENT) AND RELATIVE ERRORS

The transfer from an absolute (apparent) error to the relative error of an approximate number and the converse is possible under the following conditions:

1. If the first significant digit z of the absolute (apparent) error is given, then:

a. If the number of significant digits to which the approximate number a is correct is k, then

$$\wp'_a < \frac{z}{10^k} < \frac{1}{10^{k-1}}.$$

b. If the number of significant digits to which the approximate number a is correct is not stated (this may happen in the case of a final numerical result of numerical computations), then the number of significant digits should be determined by means of the absolute (apparent) error. For example, a result 4542.65 is obtained with a final absolute (apparent) error -1.25. The number of significant digits is then 3, and the relative error (here we can use the upper limit of the relative error) is

$$|\wp_a| < \frac{1}{4 \cdot 10^2} = 0.0025,$$

$$\wp'_a < \frac{1}{10^2}.$$

2. If the relative error $\wp_a$ is given, then

a. If the number of significant digits of the approximate number a, whose first significant digit on the extreme left is x_1, is k, then

$$|\wp_a| < \frac{z}{10^k x_1},$$

and

$$z > 10^k x_1 |\wp_a|.$$

b. If the number of significant digits to which the approximate number a (whose first significant digit on the extreme left is x_1) is correct, is not given, then the number of significant digits should be determined by inspection of the apparent and the relative errors. For this purpose the upper limit of the relative error is very important. Use should be made then of Table 4 on page 127. For example, a result 6,473.45 is obtained. The relative error of the final result is +0.00045. We have then

$$0.0001 < 0.00045 < 0.001.$$

The first significant digit of the apparent error is 3. It follows then that the approximate number is correct to 3 significant digits, and the final numerical result should be written as $6{,}473 + 3$, or $6{,}480 - 4$.

XVI. The Order of the Errors

We shall introduce an important criterion, the idea of *the order of the errors.* This criterion will be used in the process of the development of the theory of Approximate Computation.

The limits of the absolute error of an approximate number a, which is given as correct to k significant digits, are

$$-0.5 \cdot 10^{p+n-k+1} \leqslant \alpha \leqslant 0.5 \cdot 10^{p+n-k+1},$$

or

$$|\alpha_1| \leqq 0.5 \cdot 10^{p+n-k+1}.$$

Then, all absolute errors having upper limits such that their significant digits on the left are located in the same place of significance belong to the same order. If we have two absolute errors α and β such that

$$\frac{|\alpha_1|}{|\beta_1|} \leqq 1,$$

then the absolute errors α and β are said to belong to the same order.

If we have two absolute errors α and β such that

$$\frac{|\alpha_1|}{|\beta_1|} \underset{\sim}{=} 10^{-r}, \qquad (r = 1, 2, 3, \cdots)$$

then the absolute error β is said to be of the rth order in comparison with the absolute error α, i.e., β is an absolute error of an order higher than the order of the absolute error α. It follows then that the absolute error β is greater than the absolute error α. The significant digit on the left of α is located to the right of the significant digit on the left of β. From the point of view of measurement (in terms of the same units) the lower the order of the absolute error, the more precise is the measurement.

Another method of comparison of absolute error is given by $|\beta| - |\alpha| > 0$. It cannot be inferred, however, from this relationship whether β is of a higher order than α, or to what extent.

The upper limit of the relative error of an approximate number a, which is given as correct to k significant digits, is given by

$$\wp'_a < \frac{1}{10^{k-1}}.$$

Then, all relative errors having upper limits of the same magnitude are said to belong to the same order. If we have two relative errors $\wp_a$ and $\wp_b$ such that

$$\frac{\wp'_a}{\wp'_b} \underset{\sim}{=} 1,$$

then the relative errors $\wp_a$ and $\wp_b$ are said to be of the same order. If we have two relative errors $\wp_a$ and $\wp_b$ such that

$$\frac{\wp'_a}{\wp'_b} \underset{\sim}{=} 10^{-s}, \quad (s = 1, 2, 3, \cdots)$$

then the relative error $\wp_b$ is said to be of the sth order in comparison with the relative error $\wp_a$; that is, $\wp_b$ is a relative error of an order greater, or higher than the order of the relative error $\wp_a$. It follows then that

$$|\wp_a| < |\wp_b|.$$

From the point of view of the process of measurement (in terms of the same units as well as in terms of different units), the higher the order of the relative error, the less accurate is the measurement. In other words, the smaller the percentage error the more accurate is the measurement.

XVII. The Representation of Approximate Numbers

The representation of approximate numbers is effected by approximating exact numbers

1. within a given number of decimal places;
2. within a given number of significant digits;
3. within a stated apparent error;
4. within a stated relative error.

From the point of view of computation the approximation of a number within a stated number of decimal places is the least advantageous method. Sooner or later in the process of computation, if an approximation is stated to a certain number of decimal places, it becomes necessary to secure additional information relating to the accuracy of the measurement or to the precision of the measurement, and their respective errors must be determined. It follows then that in all processes of Approximate Computation *the use of the decimal places approximativeness should by all means be discouraged.*[7]

The following symbolism will be used hereafter in the process of the development of the theory of Approximate Computation.

$A, B, C, \cdots\cdots$, etc. will denote exact values

$A_1, B_1, C_1, \cdots\cdots$, etc. will denote upper limits of the true values

$A_2, B_2, C_2, \cdots\cdots$, etc. will denote lower limits of the true values

$a, b, c, \cdots\cdots$, etc. will denote approximate values of $A, B, C, \cdots$, etc.

$\alpha, \beta, \gamma, \cdots\cdots$, etc. will denote absolute errors of $a, b, c, \cdots$, etc. respectively.

$\alpha_1, \beta_1, \gamma_1, \cdots\cdots$, etc. will denote upper limits of the absolute errors $\alpha, \beta. \gamma$, etc. respectively.

$\wp_a, \wp_b, \wp_c, \cdots\cdots$, etc. will denote relative errors of a, b, c, etc. respectively.

$\wp'_a, \wp'_b, \wp'_c, \cdots\cdots$, etc. will denote upper limits of the relative errors $\wp_a, \wp_b, \wp_c$, etc. respectively.

---

[7] British Civil Service Commission; Syllabus of the Examinations; (no year); "Everything must be thought of in terms of significant figures, *never in decimal places.*"

CHAPTER XV

THE METHODS OF APPROXIMATE COMPUTATION

1. The Two Types of Problems in Approximate Computation

Two principal types of problems arise in the process of Approximate Computation, that is, in the process of application of formulas (mathematically exact, mathematically approximate, and scientific) to numerical data (exact or approximate).

1. The nature of the numerical data is definitely given, and it is required to apply formulas and to perform numerical computations in such a manner that an *optimum* approximativeness is obtained, after all the stated computations are performed.

In this case, if the numerical data are approximate, the nature of the approximation is given, and the optimum approximativeness implies that the result of the performed numerical computations is obtained without any unnecessary loss in the approximation while the process of computation takes place. It should be understood, however, that this implies that the maximum error within the above limitations should be obtained. Moreover, it implies that the approximativeness of the numerical result that is to be obtained depends entirely on the nature of the formulas and on the nature and the extent of the approximativeness of the numerical data to which these formulas are to be applied.

2. The nature of the numerical data is definitely given, and it is required to apply formulas and to perform numerical computations in such a manner that a result with a definitely *predetermined* approximativeness is obtained, after all the stated computations are performed.

In this case, if the numerical data are approximate, the nature of the approximation is given. The required approximativeness of the numerical result of the computations is also stated. The requirement of the problem is that the numerical data (as well as the formulas) fit into the problem. It will be shown later that *at no time* can a result be obtained which is more accurate or more precise, that is, which approximates better, than the least accuracy or the least precision of the approximate numerical data (except in certain cases of the extraction of roots).[1] Certain formulas, especially scientific (see Chapter XIV, sec. [illegible] c), are limited in their approximativeness, and since it is a prerequisite that the formulas fit into the problem, a preliminary examination of the formulas must be made, in order to determine the extent to which the particular formula, or formulas, may be used.

If approximate mathematical formulas are used, and such formulas are usually given in a form that provides for any desired degree of approximativeness, the extent of the formula must be definitely predetermined before the application of the formula to the numerical data takes place. In other words, the effect produced by the formula on the final result of the numerical computation must be known to the computer before he begins to perform the numerical processes. However, the evaluation of this effect may take place after the numerical computation.

A purely mathematical formula, on the other hand, may lead to the preliminary (and absolutely necessary) examination of the numerical data and of the formula itself. For example, there is no sense in the problem (unless mechanical proficiency is the aim): "Find the roots of the quadratic equation

$$2x^2 + 13x - 25 = 0$$

correct to 4 significant digits, or correct to thousandths." It might happen that this particular equation arises in connection with a money problem, and hundredths might be allowable and tolerable. It might happen that this equation is connected with a precision measurement, and even thousandths might be tolerable. But then the data of the problem would supply the information necessary for the guidance of the computer.

In other words, Approximate Computation is not strictly a mechanical process. Approximate Computation requires that data-analysis be its guiding principle. Mechanical processes aiming beyond any reason at absurd astronomical numbers must be banished from the numerical processes of any type of computation. On the other hand, a numerical process and its results have meaning, and this meaning should be interpreted in conformity with the mathematical principles that are at the basis of the theory of Approximate Computation. If rules for the elementary stages of mathematical education are stated, these rules should not be treated or accepted as mechanical; at any stage of Approximate Computation the data on hand must be used critically, and the understanding of the processes is of prime importance.

---

[1] This is directly in conformity with the cumulative property of errors. C. Fr. Gauss in his Theoria Motus formulated this principle. See sec. IV of Chapter XXII.

II. THE APPROXIMATIVENESS OF NUMERICAL RESULTS

Approximate numbers are given (we shall omit the decimal places approximativeness from the general treatment) as correct—

1. within a stated absolute (apparent) error;
2. within a stated relative error;
3. to a stated number of significant digits.

Numerical results of applications of formulas to the numerical data may be required to be stated:

1. In case the numerical data are given as correct within a stated apparent error—
    - a. with an apparent error approximativeness;
    - b. with a relative error approximativeness;
    - c. with a significant digits approximativeness.
2. In case the numerical data are given as correct within a stated relative error—
    - a. with an apparent error approximativeness;
    - b. with a relative error approximativeness;
    - c. with a significant digits approximativeness.
3. In case the numerical data are given as correct to a stated number of significant digits—
    - a. with an apparent error approximativeness;
    - b. with a relative error approximativeness;
    - c. with a significant digits approximativeness.

Thus, in total, for each type of the two general problems in Approximate Computation there are nine distinct problems, or three classes with three problems in each class.

These three classes are not completely distinct, however. There is a relation, as was shown (secs. XIII and XV of Chapter XIV), between the absolute (apparent) and relative errors, and between the number of significant digits of an approximate number and the upper limits of its two types of errors. At any stage of Approximate Computation it is, therefore, permissible to pass from one criterion for approximativeness to another, provided such a step is of advantage to the final result of the processes of Approximate Computation.

It is, however, very important that in any one given problem which involves Approximate Computation, all the stated numerical data be stated uniformly with one type of approximativeness for the entire set of data. If the data are given with two or more types of approximativeness the first thing that should be done by the computer is to transform the data so that their approximativeness is of one type, preferably the most advantageous.

CHAPTER XVI

APPROXIMATE COMPUTATION OF THE FIRST TYPE

I. General Remarks

In the problem of Approximate Computation of the first type it is required to perform numerical computations with the approximate numerical data in such a manner that an *optimum* approximativeness is obtained. In other words, this problem may be restated as follows:

> *A set of approximate numerical data with which certain numerical computations are to be performed is given. It is required to determine the extent of the approximativeness of the final numerical results after all the computations have been completed.*

The numerical operations that are to be considered and examined are as follows:

1. Addition.
2. Subtraction.
3. Multiplication.
4. Raising to a power.
5. Division.
6. Extraction of roots.
7. Any combination of these six operations.

The analysis of the various operations will be developed mostly with the use of absolute errors. This will make possible the presentation of this analysis in the most general form. The entire work hereafter will be carried on with the absolute values of the absolute errors and relative errors whenever necessary.

The attention of the reader is called to the formulation of the procedure in Approximate Computation given by Gauss in his Theoria Motus Corporum Coelestium,[1] in which he states that in all numerical processes the errors of approximate numbers should be affected by the same sign. In other words, in all numerical processes it is necessary that the errors shall be of the same direction. Should some of the approximate numbers have errors with opposite signs, that is, of different directions (some positive and some negative), then the directions of the errors must be changed so that all the errors will have the same sign.[2] Should some approximate numbers have errors with the sign ±, and some errors with a positive sign (or negative sign), then the method of bounds must be employed for the processes of Approximate Computation.[3]

---

[1] See sec. IV of Chapter XXII where this portion of Theoria Motus is quoted. Also see Chapter XIII where this restriction is fully discussed.

The above remarks refer to processes of Approximate Computation of the first and second types as well, and the reader should keep them in mind while working with approximate numerical data and while evaluating the final numerical results.

II. ADDITION

Let A and B be two exact numbers whose approximations are a and b respectively. Let their respective absolute errors be α and β. Then

$$|a| \leqslant |A| + |\alpha|,$$
$$|b| \leqslant |B| + |\beta|,$$

and

$$|a + b| \leqslant |A + B| + (|\alpha| + |\beta|).$$

But

$$\operatorname*{Limit}_{\epsilon_1 \to 0} |\alpha + \epsilon_1| = |\alpha_1|,$$

and

$$\operatorname*{Limit}_{\epsilon_2 \to 0} |\beta + \epsilon_2| = |\beta_1|.$$

Hence,

$$\operatorname*{Limit}_{\substack{\epsilon_1 \to 0 \\ \epsilon_2 \to 0}} |\alpha + \beta + \epsilon_1 + \epsilon_2| \leqslant |\alpha_1| + |\beta_1|.$$

We have then: The absolute error of the sum of two approximate numbers a and b, whose respective absolute errors are α and β, may be equal to the sum of the absolute values of the two respective absolute errors. The same may be said in reference to the apparent error of the sum of two approximate numbers.

The upper limit of the absolute error of the sum of two approximate numbers a and b, whose respective absolute errors are α and β, may be equal to the sum of the absolute values of the upper limits of the respective absolute errors.

The above rule may be extended to the sum of more than two approximate numbers. We have then

---

[2] See Appendix I, "Rounding Off Digits."

[3] See Chapter XI.

$$\left|\sum^{n} a\right| \leqslant \left|\sum^{n} A\right| + \sum^{n} |\alpha|,$$

and

$$\underset{\substack{\epsilon_i \to 0 \\ (i = 1, 2, 3, \ldots, n)}}{\text{Limit}} \left|\sum^{n} \alpha + \sum^{n} \epsilon_i\right| \leqslant \sum^{n} |\alpha_1|.$$

$$\sum_{i=1}^{n} \epsilon_i \to 0$$

We have then: The absolute error of the sum of several approximate numbers a, b, c, . . . , etc., whose respective absolute errors are α, β, γ, . . . , etc., may be equal to the sum of the absolute values of the respective absolute errors. The same is true in reference to the apparent error of a sum of several approximate numbers.

The direction of the absolute (apparent) error of the sum of approximate numbers is the same as that of the respective errors.

The upper limit of the absolute error of the sum of several approximate numbers a, b, c, . . . , etc., whose respective absolute errors are α, β, γ, . . . , etc., may be equal to the sum of the absolute values of the upper limits of the respective absolute errors.

A. THE RELATIVE ERROR

The relative errors of the two approximate numbers a, and b, whose respective absolute errors are α and β, and whose respective exact values are A and B, are

$$\wp_a = \frac{\alpha}{A} \quad \text{and} \quad \wp_b = \frac{\beta}{B}.$$

Assume that

$$|\wp_a| < |\wp_b|.$$

Then

$$\frac{|\alpha|}{|A|} < \frac{|\alpha| + |\beta|}{|A + B|} < \frac{|\beta|}{|B|},$$

or

$$|\wp_a| < |\wp_{(a+b)}| < |\wp_b|.$$

Similarly, since

$$\alpha + \epsilon_1 = \alpha_1, \ (\epsilon_1 \to 0)$$

and

$$\beta + \epsilon_2 = \beta_1, \ (\epsilon_2 \to 0)$$

we have

$$\frac{|\alpha_1|}{|A_2|} < \frac{|\alpha_1| + |\beta_1|}{|A_2 + B_2|} < \frac{|\beta_1|}{|B_2|},$$

and

$$\wp'_a < \wp'_{(a+b)} < \wp'_b.$$

The above can be extended to the case of a sum of more than two approximate numbers. We have then, assuming that $\wp_a$ is the smallest relative error, and $\wp_n$ is the greatest relative error, all the relative errors being taken in their absolute values,

$$\frac{|\alpha|}{|A|} < \frac{|\alpha| + |\beta| + |\gamma| + \cdots + |\nu|}{|A + B + C + \cdots + N|} < \frac{|\nu|}{|N|},$$

or

$$|\wp_a| < \left|\wp_{\sum_a^n}\right| < |\wp_n|.$$

And

$$\frac{|\alpha_1|}{|A_2|} < \frac{|\alpha_1| + |\beta_1| + |\gamma_1| + \cdots + |\nu_1|}{|A_2 + B_2 + C_2 + \cdots + N_2|} < \frac{|\nu_1|}{|N_2|},$$

$$\wp'_a < \wp'_{\sum_a^n} < \wp'_n.$$

We have then: The absolute value of the relative error of the sum of several approximate numbers a, b, c, . . . , etc., whose respective relative errors are $\wp_a$, $\wp_b$, $\wp_c$, . . . , etc., is greater than the absolute value of the smallest relative error and smaller than the absolute value of the greatest relative error of the respective approximate numbers. The same relationship exists in reference to the upper limits of the relative error of the sum and the upper limits of the respective approximate numbers.

B. THE ABSOLUTE AND RELATIVE ERRORS

Consider the following sum

$$347.263 + 0.0004$$

and

$$25.42 + 0.003.$$

The first approximate number is correct to thousandths; the second approximate number is correct to hundredths only. The sum of these two approximate numbers, if they were added, would be

$$372.683 + 0.0034.$$

The resultant sum cannot be correct to thousandths any more, because there is an integral error in that unit. Indeed, two approximate numbers which represent measurements of different prec[illegible] can be added with the certainty of preserving *at the best the le[illegible]ecision, but never the greater precision.*

The above conclusion can be verified by examining two approximate numbers a and b, whose absolute errors are α and β respectively. Let $|\alpha_1| < |\beta_1|$, and

$$\frac{|\alpha_1|}{|\beta_1|} \lessapprox 10^{-r}, \quad (r = 1, 2, 3, \cdots).$$

The absolute error of the sum is $|\alpha| + |\beta|$, and its upper limit is

$$|\alpha_1 + \beta_1| \leqslant |\alpha_1| + |\beta_1|.$$

But

$$\frac{|\alpha_1 + \beta_1|}{|\beta_1|} \leqslant \frac{|\alpha_1|}{|\beta_1|} + \frac{|\beta_1|}{|\beta_1|} \lessapprox 1 + 10^{-r}, \quad (r = 1, 2, \cdots).$$

Hence the absolute error of the resultant sum of the two approximate numbers a and b is at the most of the rth order in comparison with the order of the absolute error α of the more precise approximate number a. It follows then that the precision of the resultant sum of the two approximate numbers a and b is controlled by the less precise approximate number b, and the absolute (apparent) error of the lower order exerts no influence at all on the order of the absolute error of the resultant sum.

In order to obtain the correct (*precise, but not accurate*) resultant sum, the more precise approximate number should be graded down to conform with the precision of the less precise approximate number, and then the process of addition should be performed. Thus, the two approximate numbers given above should be added as follows:

$$\begin{array}{r} 347.26 + 0.004 \\ 25.42 + 0.003 \\ \hline 372.68 + 0.007 \\ \hline 372.69 - 0.003 \end{array}$$

It is quite obvious that the process of "grading down" the more precise approximate number results in a somewhat larger apparent error as well as a larger relative error of the resultant sum of the two approximate numbers.

C. THE RULE FOR ADDITION

The addition of two or more approximate numbers is performed in accordance with the following rules.

1. If two approximate numbers are to be added, they should have apparent errors of the same order.

2. The apparent error of the sum of two approximate numbers may be equal to the sum of the respective apparent errors which must be

of the same direction, and the direction of the apparent error of the sum is the same as the direction of the respective apparent errors.

3. If two approximate numbers are given such that the apparent error of one is either positive or negative, and the apparent error of the second has the sign ±, then the addition should be performed by means of bounds.[4]

4. If more than two approximate numbers are to be added, and the given approximate numbers are of different precisions, then the addition is performed in parts. First add those numbers that are more precise, and follow rules 1 and 2 above. The resultant sum should be graded down to conform to the precision of the next approximate number in order of precision, and so on, until by the method of exhaustion the total sum is obtained.

5. The resultant sum should be rounded off in order that its apparent error will be of the form

$$|\alpha| \leqslant 0.5 \cdot 10^{p+n-m+1},$$

m being the number of significant digits in the resultant sum.

III. Subtraction

Let A and B be two exact numbers, and let their approximations be a and b respectively. Let their respective absolute errors be α and β. Then

$$|a| \leqslant |A| + |\alpha|,$$
$$|b| \leqslant |B| + |\beta|,$$

and

$$|a - b| \leqslant |A - B| + (|\alpha| + |\beta|).$$

Then, the upper limit of the absolute error of the difference of the two approximate numbers a and b is

$$\underset{\substack{\epsilon_1 \to 0 \\ \epsilon_2 \to 0}}{\text{Limit}} |\alpha + \beta + \epsilon_1 + \epsilon_2| \leqslant |\alpha_1| + |\beta_1|.$$

The reader will recognize the identity of the form of the above expression with the form of the expression for the upper limit of the absolute error of the sum of two approximate numbers.

A. THE RELATIVE ERROR

The relative errors of the two approximate numbers a and b whose respective exact values are A and B, and whose respective absolute errors are α and β, are

---

[4] See Chapter XI.

$$\wp_a = \frac{\alpha}{A} \quad \text{and} \quad \wp_b = \frac{\beta}{B}.$$

Assume that

$$|\wp_a| < |\wp_b|.$$

Then we have

$$\frac{|\alpha|}{|A|} < \frac{|\alpha| + |\beta|}{|A - B|} \gtreqless \frac{|\beta|}{|B|},$$

or

$$|\wp_a| < |\wp_{(a-b)}| \gtreqless |\wp_b|.$$

Since

$$\alpha + \epsilon_1 = \alpha_1, \ (\epsilon_1 \to 0),$$

and

$$\beta + \epsilon_2 = \alpha_1, \ (\epsilon_2 \to 0),$$

we have

$$\frac{|\alpha_1|}{|A_2|} < \frac{|\alpha_1| + |\beta_1|}{|A_2 - B_2|} \gtreqless \frac{|\beta_1|}{|B_2|},$$

and

$$\wp'_a < \wp'_{(a-b)} \gtreqless \wp'_b.$$

Should

$$|\wp_a| = |\wp_b|,$$

then we have

$$|\wp_{(a-b)}| = \frac{|A| + |B|}{|A - B|} |\wp_a| > |\wp_a|.$$

We have then: The absolute value of the relative error of the difference of two approximate numbers a and b, whose respective relative errors are $\wp_a$ and $\wp_b$, $|\wp_a| \gtreqless |\wp_b|$, is greater than the absolute value of the relative error of the minuend. The same relationship exists in the case of the upper limit of the relative error of the difference and the relative errors of the respective approximate numbers.

From the above it follows that the difference of two approximate numbers should never be expected to be as accurate as the more accurate of the two approximate numbers. For example, consider the following difference:

Approximate Number	*Apparent Error*	*Relative Error*
125.1	−0.03	−0.00024
62.4	+0.01	+0.00016
62.7	−0.04	−0.00064

B. THE ABSOLUTE AND RELATIVE ERRORS

The reader should recall the development in section II, B of this chapter. In the case of subtraction, when two approximate numbers are given such that one of them is more precise than the other, the situation is exactly the same as in addition.

Let us examine two approximate numbers a and b whose absolute errors α and β are such that

$$|\alpha_1| < |\beta_1|$$

and

$$\frac{|\alpha_1|}{|\beta_1|} \leqq 10^{-r}, \quad (r = 1, 2, 3, \cdot \cdot \cdot).$$

The absolute error of the difference of the two approximate numbers a and b is $|\alpha| + |\beta|$, and its upper limit is

$$|\alpha_1 + \beta_1| \leqslant |\alpha_1| + |\beta_1|.$$

But

$$\frac{|\alpha_1 + \beta_1|}{|\beta_1|} \leqslant \frac{|\alpha_1|}{|\beta_1|} + \frac{|\beta_1|}{|\beta_1|} \leqq 1 + 10^{-r}, (r = 1, 2, 3, \cdot \cdot \cdot).$$

Hence, the absolute error of the resultant difference of the two approximate numbers a and b is at the most of the rth order in comparison with the order of the absolute error α of the more precise approximate number a. It follows then that the precision of the resultant difference of the two approximate numbers a and b is controlled by the less precise approximate number b, and the absolute error of the lower order exerts no influence at all on the order of the absolute error of the resultant difference.

In order to obtain the most correct (*precise, but not accurate*) resultant difference, the more precise approximate number should be graded down to conform with the precision of the less precise approximate number, and then the process of subtraction should take place. For example, the difference of the two approximate numbers

$$347.263 + 0.0004$$

and

$$225.42 \quad - 0.003$$

is obtained as follows:

$$\begin{array}{ll} 347.26 & + 0.004 \\ 225.42 & - 0.003 \\ \hline 121.84 & + 0.007 \\ \hline 121.85 & - 0.003 \end{array}$$

The fact that the relative error of the difference of two approximate numbers is considerably increased as a result of such a grading down can be seen from the following example:

Approximate number	Apparent error	Relative error
347.26	+0.004	+0.0000116
225.42	−0.003	−0.0000133
121.84	+0.007	+0.0000574
121.85	−0.003	−0.0000246

C. THE RULE FOR SUBTRACTION

Subtraction of approximate numbers should be performed in accordance with the following rules.

1. The subtrahend and minuend should have apparent errors of the same order.

2. The apparent error of the difference of two approximate numbers may be equal to the sum of the respective apparent errors, which must be of opposite directions, and the direction of the apparent error of the difference is the same as the direction of the apparent error of the minuend.

3. If two approximate numbers whose difference is to be obtained are given such that the apparent error of one is either positive or negative, and the apparent error of the other has the sign ±, then the subtraction should be performed by means of bounds.[5]

4. The resultant difference should be rounded off in order that its apparent error will be of the form

$$|\alpha| \leqslant 0.5 \cdot 10^{p+n-m+1},$$

m being the number of significant digits in the resultant difference.

IV. MULTIPLICATION

Let A and B be two exact numbers, whose approximations are a and b respectively. Let their respective absolute errors be α and β. Then

$$|a| \leqslant |A| + |\alpha|,$$
$$|b| \leqslant |B| + |\beta|.$$

Then

$$|a||b| \leqslant (|A| + |\alpha|)(|B| + |\beta|),$$

or

$$|a||b| \leqslant |A||B| + |A||\beta| + |B||\alpha| + |\alpha||\beta|.$$

---

[5] See Chap. XI.

By virtue of the fact that

$$\alpha = |A - a| < \epsilon_1, \quad (\epsilon_1 \rightarrow 0),$$

and

$$\beta = |B - b| < \epsilon_2, \quad (\epsilon_2 \rightarrow 0),$$

we have

$$|\alpha||\beta| < \epsilon_1\epsilon_2, \quad (\epsilon_1 \rightarrow 0; \epsilon_2 \rightarrow 0).$$

Therefore,

$$|a||b| \leqslant |A||B| + (|A||\beta| + |B||\alpha|).$$

But

$$\operatorname*{Limit}_{\epsilon_3 \rightarrow 0} |\alpha + \epsilon_3| = |\alpha_1|,$$

and

$$\operatorname*{Limit}_{\epsilon_4 \rightarrow 0} |\beta + \epsilon_4| = |\beta_1|.$$

Hence

$$\operatorname*{Limit}_{\substack{\epsilon_3 \rightarrow 0 \\ \epsilon_4 \rightarrow 0}} |A(\beta + \epsilon_4) + B(\alpha + \epsilon_3)| \leqslant |A||\beta_1| + |B||\alpha_1|.$$

The expression for the apparent error of the product of two approximate numbers is identical in form and nature with the expression for the absolute error of the product of two approximate numbers. Since

$$0 < A < A_1 \quad \text{and} \quad 0 < B < B_1$$

we have

$$|A||\beta_1| + |B||\alpha_1| < |A_1||\beta_1| + |B_1||\alpha_1|.$$

The above development may be extended to the general case of the product of any number of approximate numbers. Let a, b, c, . . . , m be the approximate numbers, whose respective absolute errors are α, β, γ, . . . , μ, and whose respective exact values are A, B, C, . . . , M. The upper limits of the true values are A_1, B_1, C_1, . . . , M_1, and the upper limits of the absolute errors are α_1, β_1, γ_1, . . . , μ_1. Then the absolute error of the product of the several approximate numbers may be equal to

$$\sum|\alpha||B||C||D| \cdot \cdot \cdot |M|,$$

summed to as many terms as there are given approximate numbers.

The upper limits of the absolute error of the product of several approximate numbers may be equal in its absolute value to

$$\sum|\alpha_1||B||C||D| \cdot \cdot \cdot |M|,$$

summed to as many terms as there are given approximate numbers to form the product. The absolute value of the upper limit of the absolute error of the product of the several approximate numbers is always less than

$$\sum |\alpha_1||B_1||C_1||D_1| \cdots |M_1|,$$

summed to as many terms as there are given approximate numbers to form the product.

The expression for the apparent error of the product of several approximate numbers is identical in form and nature with the expression for the absolute error of the product of the several approximate numbers.

A. THE RELATIVE ERROR

The relative errors of the two approximate numbers a and b, whose respective absolute errors are α and β, and whose respective exact values are A and B, are

$$\wp_a = \frac{\alpha}{A} \quad \text{and} \quad \wp_b = \frac{\beta}{B},$$

and

$$\frac{|\alpha|}{|A|} < \frac{|A||\beta| + |B||\alpha|}{|AB|} > \frac{|\beta|}{|B|},$$

or

$$\frac{|\alpha|}{|A|} < \frac{|\alpha|}{|A|} + \frac{|\beta|}{|B|} > \frac{|\beta|}{|B|},$$

that is,

$$|\wp_a| < |\wp_{(ab)}| > |\wp_b|.$$

Similarly, since

$$\alpha + \epsilon_1 = \alpha_1, \quad (\epsilon_1 \rightarrow 0),$$

and

$$\beta + \epsilon_2 = \beta_1, \quad (\epsilon_2 \rightarrow 0),$$

we have

$$\frac{|\alpha_1|}{|A_1|} < \frac{|A_1||\beta_1| + |B_1||\alpha_1|}{|A_1B_1|} = \frac{|\alpha_1|}{|A_1|} + \frac{|\beta_1|}{|B_1|} > \frac{|\beta_1|}{|B_1|},$$

and then

$$\frac{|\alpha_1|}{|A_2|} < \frac{|\alpha_1|}{|A_2|} + \frac{|\beta_1|}{|B_2|} > \frac{|\beta_1|}{|B_2|}.$$

Finally, we have

$$\wp'_a \leqslant \wp'_{(ab)} \geqslant \wp'_b.$$

We have then: The absolute value of the relative error of the product of two approximate numbers a and b whose respective relative errors are $\wp_a$ and $\wp_b$ is greater than the absolute value of the relative errors of either of the approximate numbers. The absolute value of the relative error of the product of two approximate numbers may be equal to the sum of the absolute values of the relative errors of the respective approximate numbers, that is,

$$|\wp_{(ab)}| \leqslant |\wp_a| + |\wp_b|.$$

Relative to the upper limit of the relative error of the product of two approximate numbers similar relationships exist, that is,

$$\wp'_{(ab)} \leqslant \wp'_a + \wp'_b.$$

The above development may be extended to the general case of the product of several approximate numbers. Let $a, b, c, \ldots, m$ be the approximate numbers whose respective relative errors are $\wp_a, \wp_b, \wp_c, \ldots, \wp_m$. We then may obtain

$$|\wp_k|_{(k=a,\, b,\, c,\, \ldots,\, m)} < |\wp_{(abc \ldots m)}|.$$

Similarly, if there are not more than nine factors, we have

$$\wp'_{k\ (k=a,\, b,\, c,\, \ldots,\, m)} \leqslant \wp'_{(abc \ldots m)}.$$

Otherwise,

$$\wp'_{k\ (k=a,\, b,\, c,\, \ldots,\, m)} < \wp'_{(abc \ldots m)}.$$

Also,

$$|\wp_{(abc \ldots m)}| \leqslant \sum_{k=a}^{m} |\wp_k|,$$

and

$$\wp'_{(abc \ldots m)} \leqslant \sum_{k=a}^{m} \wp'_k.$$

From the above it follows that the product of two or more approximate numbers should never be expected to be more accurate than the least accurate of the approximate numbers that enter in this product as factors. For example, consider the following product:

Approximate number	Apparent error	Relative error
125.1	+0.03	+0.00024
62.4	+0.01	+0.00016
7,806.24	+3.13	+0.00040

According to the absolute error test the resultant product is correct to three significant digits only, and this is substantiated by the relative error test, because

$$0.0001 < 0.0004 < 0.001.$$

and the first significant digit on the extreme left of the resultant product is greater than the first significant digit on the left of the apparent error. Rounding off the number representing the resultant product, we finally obtain

$$7{,}810 - 1.$$

B. THE DIRECTION OF THE ABSOLUTE (APPARENT) AND RELATIVE ERRORS OF THE PRODUCT OF APPROXIMATE NUMBERS

If the absolute (apparent) errors of the approximate numbers which form the product are of the same direction, that is, all the errors are either positive (or negative), or all have the sign ±, then the direction of the absolute (apparent) error of the resultant product is the same as the direction of the apparent errors of the respective approximate numbers.

If the apparent errors of the two (or more, but we shall consider the case of two approximate numbers, and the generalization will follow) approximate numbers are of opposite directions, that is, one of the apparent errors is positive, and the other is negative, the direction of the apparent error of the resultant product of the two approximate numbers cannot be determined by inspection of the directions of the apparent errors of the respective approximate numbers.

It might be suggested that there could be made an exception to the general procedure that in every respective process of Approximate Computation apparent errors should be affected by the same signs, and that we use the algebraic sum of the errors. Let us consider the following example:

Approximate number	Apparent error	Relative error
22.4	−0.04	−0.001786
27.9	+0.05	+0.001792
624.96	+0.004	+0.000006

In other words, we multiplied two approximate numbers, each correct to three significant digits, and we obtained a product correct to five significant digits, and the relative error of the product is almost 300 times as small as the relative error of either of the approximate numbers whose product was obtained! A result of this type is far from being acceptable even from the common-sense point of view. The reader will find no difficulty in obtaining the product of (3.98 − 0.004) and (4.98 + 0.005) correct (!) to 6 significant digits.

The fact that the process of obtaining the algebraic sum of the errors is unsatisfactory will become more obvious if we consider the following three products:

1. $(12.52 + 0.004)$ and $(6.34 - 0.002)$;
2. $(12.53 + 0.004)$ and $(6.34 - 0.003)$;
3. $(12.52 + 0.003)$ and $(6.34 - 0.002)$.

The reader will find no difficulty in obtaining the following results:

2. $79.4 \quad - 0.03$;
3. $79.37 \quad + 0.001$;
1. $79.377 + 0.0002$.

If we perform the above multiplications in modified form, however, that is, if we use the upper limits of the true values,[6] we have:

1. $(12.53 - 0.006) \cdot (6.34 - 0.002) = 79.4 - 0.02$;
2. $(12.53 - 0.006) \cdot (6.34 - 0.003) = 79.4 - 0.03$;
3. $(12.53 - 0.007) \cdot (6.34 - 0.002) = 79.4 - 0.03$.

The following analysis will demonstrate the justification of the use of the upper limits of the true values.

Let

$$0 < A_2 < A < A_1,$$

and

$$0 < B_2 < B < B_1.$$

$$\begin{aligned} a &= A_1 - e = A_2 + e', \quad |e| \leqslant |e'| \quad \text{or} \quad |e'| \leqslant |e|; \\ b &= B_1 - f = B_2 + f', \quad |f| \leqslant |f'| \quad \text{or} \quad |f'| \leqslant |f|. \end{aligned}$$

When the directions of the apparent errors are opposite, the product ab is:

1. $|ab| \leqslant |A_1B_2| + |A_1||f'| + |B_2||e|, \quad |f'| \leqslant |f|, |e| \leqslant |e'|.$
2. $|ab| \leqslant |A_2B_1| + |A_2||f| + |B_1||e'|, \quad |e'| \leqslant |e|, |f| \leqslant |f'|.$

Let us replace A_2 by A_1 and B_2 by B_1 and by their corresponding apparent errors. We have then

$$\begin{array}{ll} 1. \quad |ab| \leqslant |A_1B_2| + |A_1||f'| + |B_2||e|, & \\ & |f'| \leqslant |f|, \\ 1.1 \quad |ab| \leqslant |A_1B_1| + |A_1||f| + |B_1||e|, & \\ \qquad |A_1||f'| + |B_2||e| \leqslant |A_1||f| + |B_1||e|. & \end{array}$$

$$\begin{array}{ll} 2. \quad |ab| \leqslant |A_2B_1| + |A_2||f| + |B_1||e'|, & \\ & |e'| \leqslant |e|, \\ 2.2 \quad |ab| \leqslant |A_1B_1| + |A_1||f| + |B_1||e|, & \\ \qquad |A_2||f| + |B_1||e'| \leqslant |A_1||f| + |B_1||e|. & \end{array}$$

---

[6] See Appendix 1 on "Rounding Off Digits."

If we refer to the expression for the upper limit of the absolute error of the product of two approximate numbers (see sec. IV above) we have that the upper limit, in its absolute value, is less than

$$|A_1||\beta_1| + |B_1||\alpha_1|.$$

To the apparent error of the product *ab* in 1.1 and 2.2 there corresponds an absolute error such that its upper limit is

$$|A_1||\beta_1| + |B_1||\alpha_1|.$$

The use of the upper limits of the true values of the approximate numbers brings the apparent error of the product of the two approximate numbers closer to the vicinity of its upper limit, and we are interested in the maximum apparent error within the limitations of the approximate data.

C. THE ABSOLUTE AND RELATIVE ERRORS

The analysis of the approximativeness of a product of two (or more) approximate numbers has before it the following problems:

1. The determination of the procedure most advantageous from the point of view of the time element and the labor element, by means of which the approximativeness of the product may be evaluated.

2. The application of the various tests—(a) the absolute error test, (b) the relative error test, and (c) the combination of these two tests—to the product of approximate numbers, in order to determine the limits of approximativeness of this product.

i. The magnitude of the absolute error

The determination of the approximativeness of the product of two approximate numbers may be effected by means of the absolute (apparent) error of this product. This method of analysis is very involved, however, and from the point of view of the time and labor elements it is prohibitive.

The process of analysis is as follows:

1. The total number of digits in the product of *ab* is determined.

2. The total number of digits in the absolute (apparent) error of this product is determined.

3. The location of the first significant digit on the left of the absolute (apparent) error of the product is determined.

4. As a result of the operation indicated in 3 the last place of significance on the right of the product *ab* is determined.

The detailed analysis involves the following cases:

$$(1)\ |A| > |B|,$$
$$(2)\ |A| = |B|,$$
$$(3)\ |A| < |B|.$$

In each case the number of significant digits of a and b must be taken into consideration. Also the orders of the error α and β must be considered. Thus the analysis becomes very long and involved.

ii. The relative error

A more elegant and direct method of analysis of the approximativeness of the product of approximate numbers is the relative error test. We shall examine the product of two approximate numbers first, and we shall consider the two cases:

(1) a and b are correct to k significant digits each.

(2) a is correct to k significant digits, and b is correct to $(k + 1)$ significant digits.

(1) Let a and b be given as correct to k significant digits each. Let

$$|\wp_a| < |\wp_b|,$$

and

$$\frac{\wp'_a}{\wp'_b} \eqsim 10^{-r}, \ (r = 0, 1).$$

Then, since

$$\wp'_{(ab)} \leqslant \wp'_a + \wp'_b,$$

$$\frac{\wp'_{(ab)}}{\wp'_b} \leqslant \frac{\wp'_a + \wp'_b}{\wp'_b} = \frac{\wp'_a}{\wp'} + \frac{\wp'_b}{\wp'_b} \eqsim 1 + 10^{-r}.$$

The comparison of the upper limit of the relative error of the product is made with the upper limit of the relative error of b because b is the less accurate approximate number of the two numbers a and b.

The above result states that the upper limit of the relative error of the product of two approximate numbers (each given as correct to the same number of significant digits, and hence the relative error of this product itself), is at least of the same order as the relative error of the less accurate number.

Let the first significant digits on the extreme left of the two approximate numbers a and b be x_1 and y_1 respectively, and

$$x_1 = 1, 2, 3, \cdots, 9; \quad y_1 = 1, 2, 3, \cdots, 9; \quad x_1 \gtreqless y_1.$$

Then

$$|\wp_a| < \frac{1}{x_1 10^{k-1}}, \quad \text{and} \quad |\wp_b| < \frac{1}{y_1 10^{k-1}}.$$

Since

$$|\wp_a| < |\wp_b|, \quad x_1 \geqslant y_1.$$

We have then

$$|\wp_{(ab)}| \leqslant \frac{1}{x_1 10^{k-1}} + \frac{1}{y_1 10^{k-1}} = \frac{x_1 + y_1}{x_1 y_1 10^{k-1}}.$$

But since

$$|\wp_{(ab)}| > |\wp_b|,$$

$$\frac{x_1 + y_1}{x_1 y_1 10^{k-1}} > \frac{1}{y_1 10^{k-1}}.$$

Also

$$\frac{x_1}{x_1 + y_1} \leqslant w10^{-1}, \quad (w = 1, 2, 3, \cdot \cdot \cdot , 9).$$

Hence,

$$|\wp_{(ab)}| \leqslant \frac{1}{wy_1 10^{k-2}}.$$

If $wy_1 < 10$, then

$$|\wp_{(ab)}| \leqslant \frac{1}{v10^{k-2}}, \quad (v = 1, 2, 3, \cdot \cdot \cdot , 9),$$

and the relative error of the product of the two approximate numbers a and b is of the order "one" in comparison with the relative error of the less accurate approximate number b, because

$$\frac{\wp'_b}{\wp'_{(ab)}} \eqsim 10^{-1},$$

If $wy_1 \geqslant 10$, then

$$|\wp_{(ab)}| \leqslant \frac{1}{u10^{k-1}}, \quad (u = 1, 2, 3, \cdot \cdot \cdot , 9),$$

and the relative error of the product of the two approximate numbers is of the same order as the relative error of the least accurate approximate number b, because

$$\frac{\wp'_b}{\wp'_{(ab)}} \eqsim 1.$$

The reader will find no difficulty in verifying the above results by considering the specific values of x_1 and y_1 and the respective apparent errors.

(2) Let a be correct to k significant digits, and b be correct to $(k + 1)$ significant digits. Then

$$|\wp_a| \geqslant |\wp_b|,$$

and

$$\frac{\wp'_b}{\wp'_a} \eqsim 10^{-r}, \quad (r = 0, 1, 2)$$

Then, since

$$\wp'_{(ab)} \leqslant \wp'_a + \wp'_b$$

$$\frac{\wp'_{(ab)}}{\wp'_a} \leqslant \frac{\wp'_a + \wp'_b}{\wp'_a} = \frac{\wp'_a}{\wp'_a} + \frac{\wp'_b}{\wp'_a} = 1 + 10^{-r}.$$

The comparison of the upper limit of the relative error of the product is made with the upper limit of the relative error of a because a is the less accurate approximate number of the two numbers a and b.

The above result states that the upper limit of the relative error of the product of two approximate numbers, one given as correct to k significant digits, and the other given as correct to $(k + 1)$ significant digits, and hence the relative error of this product itself, is at least of the same order as the relative error of the less accurate number.

Let the first significant digits on the extreme left of the two approximate numbers a and b be x_1 and y_1 respectively, and

$$x_1 = 1, 2, 3, \cdots, 9; \quad y_1 = 1, 2, 3, \cdots, 9; x_1 \gtreqless y_1.$$

Then

$$|\wp_a| < \frac{1}{x_1 10^{k-1}}, \quad \text{and} \quad |\wp_b| < \frac{1}{y_1 10^k}.$$

Since

$$|\wp_a| \geqslant |\wp_b|, x_1 \gtreqless y_1.$$

We have then

$$|\wp_{(ab)}| \leqslant \frac{1}{x_1 10^{k-1}} + \frac{1}{y_1 10^k} = \frac{x_1 + 10y_1}{x_1 y_1 10^k}.$$

But

$$\frac{x_1 + 10y_1}{x_1 y_1 10^k} > \frac{1}{x_1 10^{k-1}}.$$

Also

$$\frac{x_1 y_1}{x_1 + 10y_1} \leqslant w10^{-1}, \quad (w = 1, 2, 3, \cdots, 9).$$

Hence,

$$|\wp_{(ab)}| \leqslant \frac{1}{w10^{k-1}},$$

and therefore,

$$\frac{\wp'_a}{\wp'_{(ab)}} \underset{\sim}{=} 1,$$

that is, the relative error of the product of the two approximate numbers a and b is of the same order as the greatest (in absolute value) relative error $\wp_a$. The reader will find no difficulty in verifying the above result by considering the specific values of x_1 and y_1 and the respective apparent errors.

iii. The relative error of a product of three approximate numbers

The accuracy of a product of three approximate numbers a, b, and c could be determined by evaluating the approximativeness of the product of any two of these three approximate numbers, that is, (ab), (ac), or (bc), and then considering the relative errors of the partial product and of the product of this partial product with the remaining approximate number. This process, however, is disadvantageous for two reasons:

1. The relative error of the partial product may be of a higher order than the order of the relative error of the constituent factors of this partial product, or at the most it may be of the same order as the order of the relative error of the less accurate factor of this partial product. Then another product must be obtained, and by so doing we run the risk of, at the most, lowering the accuracy of the product.
2. In practice, multiplication of more than two factors is performed by means of logarithms, and this last process does not necessitate the use of partial products.

In order to avoid unnecessary loss in accuracy in the process of multiplication of more than two approximate numbers, it is *decidedly advantageous* to consider the product of all the approximate numbers as a simultaneous product, and if partial products are formed, the question of the accuracy of such partial products should be entirely ruled out. The final resultant product will thus be obtained as the most accurate within the limitations controlled by the accuracies of the respective constituent factors.

In the process of analysis of the approximativeness of the product of three approximate numbers we shall consider the following three cases:

(1) When a, b, and c are all given as correct to k significant digits each.

(2) When a and b are given as correct to k significant digits each, and c is given as correct to $(k+1)$ significant digits.

(3) When a is given as correct to k significant digits, and b and c are given as correct to $(k+1)$ significant digits each.

(1) Let a, b, and c be given as correct to k significant digits each.

Then

$$|\wp_a| \leqslant |\wp_b| < |\wp_c|,$$

and

$$\wp'_{(abc)} \leqslant \wp'_a + \wp'_b + \wp'_c.$$

Then

$$\frac{\wp'_{(abc)}}{\wp'_c} > \frac{\wp'_a + \wp'_b + \wp'_c}{\wp'_c} = \frac{\wp'_a}{\wp'_c} + \frac{\wp'_b}{\wp'_c} + \frac{\wp'_c}{\wp'_c} \eqsim 10^{-r} + 10^{-r} + 1,$$

$$(r = 0, 1).$$

Let the first significant digits on the extreme left of a, b, and c be x_1, y_1, and z_1 respectively.

$$\begin{aligned} x_1 &= 1, 2, 3, \cdots, 9; \\ y_1 &= 1, 2, 3, \cdots, 9; \qquad x_1 \gtreqless y_1 \gtreqless z_1. \\ z_1 &= 1, 2, 3, \cdots, 9; \end{aligned}$$

Then

$$|\wp_a| < \frac{1}{x_1 10^{k-1}}, \quad |\wp_b| < \frac{1}{y_1 10^{k-1}}, \quad \text{and} \quad |\wp_c| < \frac{1}{z_1 10^{k-1}}.$$

Since

$$|\wp_a| \leqslant |\wp_b| < |\wp_c|, \quad x_1 \gtreqless y_1 \gtreqless z_1.$$

We have then

$$|\wp_{(abc)}| \leqslant \frac{1}{x_1 10^{k-1}} + \frac{1}{y_1 10^{k-1}} + \frac{1}{z_1 10^{k-1}} = \frac{x_1y_1 + x_1z_1 + y_1z_1}{x_1y_1z_1 10^{k-1}}.$$

Since

$$|\wp_{(abc)}| > |\wp_b| \quad \text{and} \quad |\wp_{(abc)}| > |\wp_c|,$$

$$\frac{x_1y_1 + x_1z_1 + y_1z_1}{x_1y_1z_1 10^{k-1}} > \frac{1}{y_1 10^{k-1}},$$

and

$$\frac{x_1y_1 + x_1z_1 + y_1z_1}{x_1y_1z_1 10^{k-1}} > \frac{1}{z_1 10^{k-1}}.$$

Also

$$\frac{x_1y_1}{x_1y_1 + x_1z_1 + y_1z_1} \leqslant v10^{-1}, \quad (v = 1, 2, 3, \cdots, 9),$$

and

$$\frac{x_1z_1}{x_1y_1 + x_1z_1 + y_1z_1} \leqslant w10^{-1}, \quad (w = 1, 2, 3, \cdots, 9).$$

Hence

$$|\wp_{(abc)}| \leqslant \frac{1}{wy_1 10^{k-2}} < \frac{1}{u10^{k-2}},$$

$$|\wp_{(abc)}| \leqslant \frac{1}{vz_1 10^{k-2}} < \frac{1}{t10^{k-2}}.$$

This shows that the product of the three approximate numbers a, b, and c, under the above stated conditions, has a relative error $\wp_{(abc)}$ which is of the order "zero" or "one" in comparison with the relative error of the least accurate approximate number c, all depending on the fact whether the products wy_1 and vz_1 are simultaneously $\geqslant 10$ or are < 10. The reader will find no difficulty in verifying the above result by considering the specific values of x_1, y_1, and z_1 as well as the apparent errors of the respective approximate numbers.

(2) Let a and b be given as correct to k significant digits each, and c be given as correct to $(k + 1)$ significant digits. Then

$$|\wp_a| \geqslant |\wp_b| > |\wp_c|,$$

and

$$\wp'_{(abc)} \geqslant \wp'_a + \wp'_b + \wp'_c.$$

Then

$$\frac{\wp'_{(abc)}}{\wp'_a} \leqslant \frac{\wp'_a + \wp'_b + \wp'_c}{\wp'_a} = \frac{\wp'_a}{\wp'_a} + \frac{\wp'_b}{\wp'_a} + \frac{\wp'_c}{\wp'_a} \cong 1 + 10^{-r} + 10^{-s},$$

$$(r = 0, 1;\ s = 0, 1, 2).$$

Let the first significant digits on the extreme left of a, b, and c be x_1, y_1, and z_1 respectively.

$$x_1 = 1, 2, 3, \cdot \cdot \cdot, 9;$$
$$y_1 = 1, 2, 3, \cdot \cdot \cdot, 9; \qquad x_1 \gtreqless y_1 \gtreqless z_1.$$
$$z_1 = 1, 2, 3, \cdot \cdot \cdot, 9.$$

Then

$$|\wp_a| < \frac{1}{x_1 10^{k-1}}, \quad |\wp_b| < \frac{1}{y_1 10^{k-1}}, \quad \text{and} \quad |\wp_c| < \frac{1}{z_1 10^k}.$$

We have then

$$|\wp_{(abc)}| \leqslant \frac{1}{x_1 10^{k-1}} + \frac{1}{y_1 10^{k-1}} + \frac{1}{z_1 10^k},$$

$$|\wp_{(abc)}| \leqslant \frac{x_1y_1 + 10x_1z_1 + 10y_1z_1}{x_1y_1z_1 10^k} > \frac{1}{x_1 10^{k-1}}.$$

But

$$\frac{x_1y_1z_1}{x_1y_1 + 10x_1z_1 + 10y_1z_1} \leqslant w10^{-1}, \quad (w = 1, 2, \cdot \cdot \cdot, 9).$$

Hence,

$$|\wp_{(abc)}| \leqslant \frac{1}{w10^{k-1}}.$$

This shows that the relative error $\wp_{(abc)}$ is of the order "zero" in comparison with the relative error of the least accurate approximate number a. The reader will find no difficulty in verifying the above result by considering the specific values of x_1, y_1, and z_1 as well as the apparent errors of the respective approximate numbers.

(3) Let a be given as correct to k significant digits, and b and c be given as correct to $(k + 1)$ significant digits each.
Then

$$|\wp_a| > |\wp_b| \geqq |\wp_c|,$$

and

$$\wp'_{(abc)} \leqslant \wp'_a + \wp'_b + \wp'_c.$$

Then

$$\frac{\wp'_{(abc)}}{\wp'_a} \leqslant \frac{\wp'_a + \wp'_b + \wp'_c}{\wp'_a} = \frac{\wp'_a}{\wp'_a} + \frac{\wp'_b}{\wp'_a} + \frac{\wp'_c}{\wp'_a} \leqq 1 + 10^{-r} + 10^{-r},$$

$$(r = 0, 1, 2).$$

Let the first significant digits on the extreme left of a, b, and c be x_1, y_1, and z_1 respectively.

$$\begin{aligned} x_1 &= 1, 2, 3, \cdots, 9; \\ y_1 &= 1, 2, 3, \cdots, 9; \qquad x_1 \gtreqless y_1 \gtreqless z_1. \\ z_1 &= 1, 2, 3, \cdots, 9. \end{aligned}$$

Then

$$|\wp_a| < \frac{1}{x_1 10^{k-1}}, \quad |\wp_b| < \frac{1}{y_1 10^k}, \quad \text{and} \quad |\wp_c| < \frac{1}{z_1 10^k}.$$

We have then

$$|\wp_{(abc)}| \leqslant \frac{1}{x_1 10^{k-1}} + \frac{1}{y_1 10^k} + \frac{1}{z_1 10^k},$$

$$|\wp_{(abc)}| \leqslant \frac{x_1y_1 + x_1z_1 + 10y_1z_1}{x_1y_1z_1 10^k} > \frac{1}{x_1 10^{k-1}}.$$

But

$$\frac{x_1y_1z_1}{x_1y_1 + x_1z_1 + 10y_1z_1} \leqslant w10^{-1}, \quad (w = 1, 2, \cdots, 9)$$

Hence,

$$|\wp_{(abc)}| \leqslant \frac{1}{w10^{k-1}}.$$

This shows that the relative error $\wp_{(abc)}$ is of the order "zero" in comparison with the relative error of the least accurate approximate number a. The reader will find no difficulty in verifying the above result by considering the specific values of x_1, y_1, and z_1 as well as the apparent errors of the respective approximate numbers.

iv. The relative error of the product of n *approximate numbers*

The above treatment may be generalized to the case of the product of n approximate numbers. The relative error of such a product is

$$|\wp_{(abc\,\cdots\,n)}| \leqslant \sum_{k=a}^{n} |\wp_k|.$$

Since the method of analysis of this case is identical with the method that was employed in the preceding sections, we shall omit the details, and we shall state the final results only.

Assuming that we have a product of n approximate numbers, all given as correct to k significant digits, and that there is *one* approximate number, for instance, m, whose relative error is the greatest in its absolute value, this relative error being $\wp_m$, we have:

$$|\wp_{(abc\,\cdots\,n)}| \leqslant \frac{1}{x_{1a}10^{k-1}} + \frac{1}{x_{1b}10^{k-1}} + \cdots + \frac{1}{x_{1n}10^{k-1}},$$

where $x_{1a}, x_{1b}, \ldots, x_{1n}$ are the first significant digits on the extreme left of the respective approximate numbers $a, b, c, \ldots, n$.

Then

$$|\wp_{(abc\,\cdots\,n)}| \leqslant \frac{\sum \prod^{n-1} x_{1i}}{x_{1m}10^{k-1}\prod^{n-1} x_{1j}} > \frac{1}{x_{1m}10^{k-1}},^{*}$$

$$i = a, b, c, \cdots, n;$$
$$j = a, b, c, \cdots, l, n. \text{ (m is omitted).}$$

Let

$$n = 10r + t,$$
$$(r = 1, 2, 3, \cdots;$$
$$t = 1, 2, 3, \cdots, 9).$$

Then

$$\frac{\prod^{n-1} x_{1j}}{\sum \prod^{n-1} x_{1i}} \leqslant v10^{-(r+1)}, \quad (v = 1, 2, \cdots, 9).$$

Hence,

$$|\wp_{(abc\,\cdots\,n)}| \leqslant \frac{1}{vx_{1m}10^{k-(r+1)-1}} < \frac{1}{x_{1m}10^{k-r-2}}.$$

Finally, we have

$$\frac{\wp'_m}{\wp'_{(abc\,\cdots\,m)}} \underset{\sim}{=} 10^{-(r+1)}.$$

---

* The "summation" is performed to n terms.

The product of, at the most, *nine* approximate numbers will thus have a relative error which will be of the order "zero" or of the order "one" in comparison with the relative error of the least accurate factor. The product of *ten* up to and including 99 approximate numbers will have a relative error which will be of the order "one" or of the order "two" in comparison with the relative error of the least accurate factor.

D. SIGNIFICANT DIGITS

The determination of the approximativeness of the resultant product of several approximate numbers as to the number of significant digits, to which the product is correct, may be accomplished by means of the combined absolute (apparent) and relative error test. The magnitude of the absolute (apparent) error may be determined according to the procedure outlined on page 152. This process is, however, lengthy, and it is more advantageous to use the formula given on page 130, that is,

$$z > 10^t y_1 |\varphi_{(abc \cdots n)}|.$$

Although the actual magnitude of the apparent error of the resultant product of several approximate numbers is very important from the point of view of (1) checking the result, and (2) stating the final result with its apparent error, the detailed process of multiplication should emphasize the use of relative errors.

The determination of the number of significant digits to which the resultant product of several approximate numbers is correct is essential in the process of evaluation of the approximativeness of the resultant product. This determination of the number of significant digits performs another important function which is very useful in the case of a product of two approximate numbers. Should the number of significant digits of a product of two approximate numbers be known in advance, the computer may employ the so-called contracted multiplication, and thus save time and labor.

The result obtained in section IV, C, *ii*, above shows that the product of two approximate numbers has a relative error which is either of the order "zero" or of the order "one" in comparison with the relative error of the least accurate factor. This means that the number of significant digits in the product is either the same as or one less than in the less accurate factor. Let us obtain the product of two approximate numbers, both given as correct to the same number of significant digits:

$$a = 10^p(10^k x_0 + 10^{k-1}x_1 + 10^{k-2}x_2 + \cdots + 10x_{k-1} + x_k + 10^{-1}u),$$
$$b = 10^q(10^k y_0 + 10^{k-1}y_1 + 10^{k-2}y_2 + \cdots + 10y_{k-1} + y_k + 10^{-1}v),$$

where

$$\begin{array}{l} p \text{ is an integer} \gtreqless 0, \\ q \text{ is an integer} \gtreqless 0, \end{array} \qquad |p| \gtreqless |q|$$

$$x_0 \neq 0; \quad x_i = 0, 1, 2, \cdots, 9; \quad i = 1, 2, 3, \cdots, k;$$
$$y_0 \neq 0; \quad y_j = 0, 1, 2, \cdots, 9; \quad j = 1, 2, 3, \cdots, k.$$

The product 10^{p+q} may be disregarded, because it is not important in the analysis of the product ab. The product ab is then as follows:

$$\begin{array}{r|l} 10^{2k}x_0y_0 + 10^{2k-1}x_1y_0 + \cdots + 10^{k+1}x_{k-1}y_0 + 10^k x_ky_0 & +10^{k-1}y_0u \\ 10^{2k-1}x_0y_1 + \cdots + 10^{k+1}x_{k-2}y_1 + 10^k x_{k-1}y_1 & +10^{k-1}y_1x_k \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \cdots\cdots \\ \cdots\cdots\cdots\cdots\cdots\cdots & \cdots\cdots \\ & +10^{k-1}vx_0 \end{array}$$

After adding the similar terms we have a polynomial with $(k + 1)$ terms in it, to the left of the vertical line.

Let us examine the above product for the case when $u = 5$ and $v = 5$. This case is very important, because when the apparent error is not stated it is assumed to be ±0.5 of a unit to which the approximate number is correct.

If we assume that $u = v = 5$, then it is obvious that

$$10^{k-1}(y_0u + y_1x_k + y_2x_{k-1} + \cdots + vx_0) \geqslant 10^k,$$

and this, being the "error portion" of the above product ab, results in the fact that the product ab, although a and b are both given as correct to $(k + 1)$ significant digits, is correct to k significant digits. But x_0y_0 may be $\geqslant 10$; this is possible when

$$\begin{array}{lll} x_0 = 2, & \text{and} & y_0 \geqslant 5; \\ x_0 = 3, & \text{and} & y_0 \geqslant 4; \\ x_0 = 4, & \text{and} & y_0 \geqslant 3; \\ x_0 \geqslant 5, & \text{and} & y_0 = 2. \end{array}$$

If the product $x_0y_0 = 9$, and there is doubt whether $x_0y_0 \geqslant 10$, the product $(10x_0 + x_1)(10y_0 + y_1)$ should be examined to see whether it is $\geqslant 1{,}000$. If $x_0y_0 \geqslant 10$ or $(10x_0 + x_1)(10y_0 + y_1) \geqslant 1{,}000$, then the $(k + 1)$ significant digits are still in the product ab. Thus we have a criterion for the determination of the number of significant digits in a product of two approximate numbers when their apparent errors both have 5 as their first significant digit on the left.

Let us examine the product ab when a has $(k + 1)$ significant digits and b has $(k + 2)$ significant digits. We have then

$$a = 10^p(10^k x_0 + 10^{k-1}x_1 + \cdot \cdot \cdot + 10x_{k-1} + x_k + 10^{-1}u);$$
$$b = 10^q(10^{k+1}y_0 + 10^k y_1 + \cdot \cdot \cdot + 10^2 y_{k-1} + 10y_k + y_{k+1} + 10^{-1}v).$$

Disregarding the product 10^{p+q} we can write the product ab as follows:

$$\begin{array}{llll|l}
10^{2k+1}x_0y_0 + & \cdots\cdots & \cdots\cdots & \cdots\cdots & +10^k y_0 u \\
 & 10^{2k}x_0y_1 + & \cdots\cdots & \cdots\cdots & +10^k y_1 x_k \\
 & & 10^{2k-1}x_0y_2 + & \cdots\cdots & +10^k y_2 x_{k-1} \\
 & & & \cdots\cdots & \cdots\cdots \\
 & & & \cdots\cdots & \cdots\cdots \\
 & & & & +10^k y_{k+1} x_0
\end{array}$$

Adding the similar terms on the left of the vertical line, we obtain a polynomial with $(k + 1)$ terms in it. If we examine the expression

$$10^k(y_0u + y_1x_k + y_2x_{k-1} + \cdot \cdot \cdot + y_{k+1}x_0)$$

we find that it may be $\geqslant 10^{k+1}$. If we repeat the reasoning used in the case of the product of two approximate numbers both given as correct to $(k + 1)$ significant digits, we shall arrive at the same conclusion when one is given as correct to $(k + 1)$ and the other as correct to $(k + 2)$ significant digits. The product of the two approximate numbers is correct to k or $(k + 1)$ significant digits, and the criterion based on the product of the significant digits on the extreme left of the approximate numbers holds true in both cases.[7]

The question then arises whether there is any sense in keeping an extra significant digit in the more accurate approximate number, especially when the first significant digit of both the apparent errors is 5, and it is known that the use of an extra significant digit would at no time control the number of significant digits in the resultant product of the approximate numbers. In order to answer this question let us examine the expressions for the relative errors of the products of two approximate numbers obtained on page 153. These are:

---

[7] There might occur the so-called "borderline cases" when $x_0 = 1$, and $y_0 = 1, 2, 3, \ldots, 9$, and the product may seem to have $(k + 1)$ digits "free from error." But the apparent error of the product is greater than 0.5 of the digit in the $(k + 2)$nd place of the product, and this would reduce the number of significant digits to k. [If there were k digits "free from error" this would reduce to $(k - 1)$ significant digits.]

(1) When a and b are both given as correct to k significant digits

$$|\varphi_{(ab)}| \leqslant \frac{x_1 + y_1}{x_1 y_1 10^{k-1}}.$$

(2) When a is given as correct to k significant digits, and b is given as correct to $(k + 1)$ significant digits

$$|\varphi_{(ab)'}| \leqslant \frac{x_1 + 10y_1}{x_1 y_1 10^{k}}.$$

Then

$$|\varphi_{(ab)}| - |\varphi_{(ab)'}| \leqslant \frac{9}{y_1 10^{k}},$$

and this indicates that the relative error $|\varphi_{(ab)'}|$ is smaller in absolute value than the relative error $|\varphi_{(ab)}|$. There is then an advantage in keeping an extra significant digit in one of the factors.

If we examine the relative errors of products of three approximate numbers when

(1) a and b are given as correct to k significant digits each, and c is given as correct to $(k + 1)$ significant digits; and

(2) a is given as correct to k significant digits, and b and c are given as correct to $(k + 1)$ significant digits each; we obtain the difference (1) − (2), which is

$$\leqslant \frac{9}{y_1 10^{k}}.$$

Again, we arrive at the conclusion that it is advantageous to keep an extra significant digit in as many factors as possible.

E. RULES FOR MULTIPLICATION

Multiplication of approximate numbers should be performed in accordance with the following rules.

1. If the several approximate numbers that are to be multiplied represent measurements of different orders of accuracy, that is, if the approximate numbers do not all have the same number of significant digits, the more accurate approximate numbers, that is, the numbers having more significant digits, should be rounded off within one more significant digit than the approximate number with the least number of significant digits.

2. If the product of several approximate numbers is known to be

correct to k significant digits, then the process of multiplication should be carried to $(k + 1)$ digits.

3. The relative error of the product of several approximate numbers may be equal to the sum of the relative errors of the respective approximate numbers, and the directions of these errors must be the same.

4. If the directions of the relative (and apparent) errors of the approximate numbers are not the same, then the limits of the true values of the approximate numbers should be changed in order that the directions shall be the same, and use should be made of the upper limits of the true values. The direction of the relative error of the product is the same as the direction of the respective relative (apparent) errors of the approximate numbers.

5. If some of the approximate numbers have positive (or negative) errors, and some have errors with the sign ±, then the process of multiplication should be carried out by the method of bounds.[8]

6. If two approximate numbers are multiplied, and the approximate numbers are given as correct to significant digits with no information concerning the apparent (or relative) errors, then if the less accurate approximate number has k significant digits, the resultant product will contain k significant digits if and only if the product of the first significant digits on the extreme left of each approximate number is $\geqslant 10$, or if in doubt, the product of the first two significant digits on the extreme left of each approximate number is $\geqslant 1{,}000$. Otherwise, the product of the two approximate numbers is correct to $(k - 1)$ significant digits.

7. The apparent error of the product of several approximate numbers is determined by means of the formula[9]

$$z > y_1 10^k |\wp_{(abc \cdots n)}|.$$

8. The resultant product of several approximate numbers should be rounded off in order that the apparent error of this product will be of the form

$$|\alpha| \leqslant 0.5 \cdot 10^{p+n-m+1},$$

m being the number of significant digits in the resultant product.

---

[8] See Chap XI.

[9] See Chap. XIV, sec. XV.

CHAPTER XVII

APPROXIMATE COMPUTATION OF THE FIRST TYPE

(*Continued*)

I. RAISING TO A POWER

LET A be an exact number whose approximation is a. Let the absolute error of a be α, and the relative error of a be $\wp_a$. Then

$$|a| \leqslant |A| + |\alpha|,$$

$$|a^2| \leqslant (|A| + |\alpha|)^2,$$

or

$$|a^2| \leqslant |A|^2 + 2|A||\alpha| + |\alpha^2|.$$

By virtue of the fact that (see page 124)

$$|\alpha| = |A - a| < \epsilon_1, \quad (\epsilon_1 \rightarrow 0),$$

we have

$$|\alpha^2| < \epsilon_1^2, \quad (\epsilon_1 \rightarrow 0).$$

But

$$\underset{\alpha \to 0}{\text{Limit}}\, |\alpha + \epsilon_2| = |\alpha_1|,$$

then

$$\underset{\alpha \to 0}{\text{Limit}}\, 2|A(\alpha + \epsilon_2)| \leqslant 2|A||\alpha_1|.$$

We have then: The absolute error (apparent error) of the square of an approximate number a, whose absolute error is α, is equal to "twice the product of the absolute values of the exact value A and the absolute error α (apparent error) of the approximate number." Since

$$0 < A_2 < A < A_1,$$

we have

$$|A||\alpha_1| < |A_1||\alpha_1|,$$

which leads to the fact that the upper limit of the absolute error of the square of an approximate number a is in its absolute value always less than

$$2|A_1||\alpha_1|.$$

The above may be extended to the general case of the nth power of the approximate number a. By virtue of the results shown on page 146 we have the expression for the absolute error of the *n*th power of *a*,

$$\sum|\alpha|(|A|)^{n-1} = n|\alpha|(|A|)^{n-1},$$

and the expression for the upper limit of this absolute error is

$$\sum|\alpha_1|(|A|)^{n-1} = n|\alpha_1|(|A|)^{n-1},$$

and the upper limit of this absolute error is always less than

$$\sum|\alpha_1|(|A_1|)^{n-1} = n|\alpha_1|(|A_1|)^{n-1}.$$

The direction of the absolute (apparent) error of the *n*th power ($n = 2, 3, 4, \ldots$) of the approximate number a is the same as the direction of the absolute (apparent) error of a itself.

A. THE RELATIVE ERROR

The relative error of the approximate number a, whose exact value is A, and whose absolute error is α, is

$$\wp_a = \frac{\alpha}{A},$$

and

$$|\wp_a| < \frac{|\alpha_1|}{|A_2|}.$$

Then

$$|\wp_{(a^2)}| = 2\frac{|\alpha|}{|A|},$$

and

$$|\wp_a| < |\wp_{(a^2)}|.$$

Similarly, we have

$$\wp'_a \leqslant \wp'_{(a^2)}.$$

Also,

$$\wp'_{(a^2)} \leqslant 2\wp'_a.$$

The above may be extended to the case of the *n*th power of the approximate number a. By virtue of the results of sec. IV, A, of Chapter XVI we have

$$|\wp_a| < |\wp_{(a^n)}|, \quad (n = 1, 2, 3, \cdots),$$
$$|\wp_{(a^n)}| = n|\wp_a|,$$
$$\wp'_a \leqslant \wp'_{(a^n)}, \quad (n < 10),$$
$$\wp'_a < \wp'_{(a^n)}, \quad (n \geqslant 10),$$
$$\wp'_{(a^n)} \leqslant n\wp'_a, \quad (n < 10),$$
$$\wp'_{(a^n)} < n\wp'_a, \quad (n \geqslant 10).$$

B. THE MAGNITUDE OF THE ABSOLUTE (APPARENT) ERROR

The analysis of the approximativeness of the nth power of an approximate number a may be effected by means of the absolute and relative errors. From a practical point of view the determination of the approximativeness in terms of significant digits of a square of an approximate number presents comparatively little difficulty, and we shall concern ourselves with the phase in detail.

The absolute error of the square of an approximate number a whose absolute error is α, and whose exact value is A, is located between the limits

$$0.1 \cdot 10^{p+n-k+1} \cdot 2|A| \leqslant 2|A||\alpha| \leqslant 0.5 \cdot 10^{p+n-k+1} \cdot 2|A|,$$

and

$$2|A|10^{p+n-k} \leqslant 2|A||\alpha| \leqslant 10|A|10^{p+n-k},$$

where k is the number of significant digits.

If we take into account the first significant digits on the extreme left of A, and the fact that an absolute (apparent) error is a decimal fraction of the last unit to which an approximate number is correct, and that, therefore, the absolute (apparent) error of the square of an approximate number is also some decimal fraction of A, we may determine the number of digits in the apparent error of the square that will occupy places of significance on the right side of A, if A were multiplied by itself and no digits in A^2 were suppressed. This number of digits of the absolute (apparent) error should be subtracted from the number of digits of A^2.

The reader will find no difficulty in securing the results given in Table 5. This table gives, however, the limiting number of digits.

TABLE 5

First digit of A (on left)	Number of digits in $2 \cdot A$	Number of digits in the absolute (apparent) error coinciding with digits in A^2	
		The significant digit of the apparent error is 1	The significant digit of the apparent error is 5
<5	k	$k-1$	k
$\geqslant 5$	$k+1$	k	k

It does not give any information when the first significant digits on the left of the apparent error are either 2, or 3, or 4. From the point of view of time element the above method of evaluating approximativeness is inefficient in practical applications, however. Its value is theoretical only.

The number of digits in A^2 (if no digits were suppressed) is determined as follows: (let A have k digits)

(1) If $A \geqslant \sqrt{10^{2m+1}}$, then there are $(2k)$ digits in A^2.
(2) If $A < \sqrt{10^{2m+1}}$, then there are $(2k - 1)$ digits in A^2.

C. THE RELATIVE ERROR

Let the approximate number a be given as correct to k significant digits.

Let us denote the first significant digit on the left of the apparent (absolute) error of a by z, $z = 1, 2, 3, 4, 5$. The first significant digit on the extreme left of a is x_1. Then

$$|\wp_a| < \frac{z}{x_1 10^k},$$

and

$$|\wp_{(a^2)}| \leqslant \frac{2z}{x_1 10^k}.$$

But

$$\frac{1}{2z} = v10^{-1}, \quad (1 \leqslant v \leqslant 5).$$

Hence

$$|\wp_{(a^2)}| \leqslant \frac{1}{v x_1 10^k},$$

that is, the relative error of the square of the approximate number a is either of the order "zero" or of the order "one" in comparison with the relative error of the approximate number a itself.

The reader will find no difficulty in verifying the above result by considering the specific values of x_1 and the respective values of the apparent (absolute) error of the approximate number a.

D. THE SIGNIFICANT DIGITS OF A SQUARE OF AN APPROXIMATE NUMBER

Combining the results shown in sections I, B and I, C above we may obtain Table 6 which gives the number of significant digits in a square of an approximate number, when the approximate number a itself is given as correct to k significant digits. We shall denote by the letter M

$$\sqrt{10^{2m+1}},$$

where m is an integer $\lesseqgtr 0$.

TABLE 6

The significant digit on the left of the apparent error (z)	The approximate number a is located between the limits	The number of significant digits in a^2
1	All values	k
$2 \leqslant z \leqslant 4$	$\frac{10^m}{2z} < a \leqslant M$	$k - 1$
	$\frac{10^m}{2z} > a \geqslant M$	k
5	$a < M$	$k - 1$
	$a \geqslant M$	k

E. THE RULES FOR RAISING TO A POWER

The process of raising an approximate number to a power should be performed in accordance with the following rules.

1. The relative error of the nth power ($n = 2, 3, 4, \ldots$), of an approximate number is n times as great as the relative error of the approximate number itself.

2. The direction of the relative (and apparent) error of the nth power of an approximate number is the same as the direction of the relative (and apparent) error of the approximate number itself.

3. The number of significant digits in the square of an approximate number a given as correct to k significant digits is determined by means of Table 6.

4. If the nth power of an approximate number is known to be correct to k significant digits, then the process of raising to a power should be carried to $(k + 1)$ digits, and the $(k + 1)$st digit, counting from left to right should be used for the correction of the apparent error of the nth power.

5. The apparent error of the nth power of an approximate number is determined by means of the formula[1]

$$z > y_1 10^k |\wp_{(a^n)}|.$$

6. If an approximate number a is given as correct to k significant digits, and neither its apparent error nor its relative error is stated, then the square of the approximate number is correct to k significant digits if

$$a \geqslant \sqrt{10^{2m+1}},$$

and it is correct to $(k - 1)$ significant digits if

$$a < \sqrt{10^{2m+1}},$$

where m is an integer $\lesseqgtr 0$, and

$$10^m \cdot 3.1622776 \cdots < \sqrt{10^{2m+1}} < 10^m \cdot 3.1622777 \cdots.$$

7. The resultant nth power of an approximate number should be rounded off in order that the apparent error of this nth power is of the form

$$|\alpha| \leqslant 0.5 \cdot 10^{p+n-m+1},$$

m being the number of significant digits to which this nth power is correct.

II. DIVISION

Let A and B be two exact numbers whose approximations are a and b respectively. Let the absolute errors of the approximate numbers be α and β respectively. Then

$$\begin{aligned} |a| &\leqslant |A| + |\alpha|, \\ |b| &\leqslant |B| + |\beta|, \\ |\alpha| &= |A - a| < \epsilon_1, \quad (\epsilon_1 \to 0), \\ |\beta| &= |B - b| < \epsilon_2, \quad (\epsilon_2 \to 0). \end{aligned}$$

By virtue of the results shown on page 145 we have

$$|a||b| \leqslant |A||B| + |A||\beta| + |B||\alpha|.$$

Then

$$\frac{|a||b|}{|b|^2} \leqslant \frac{|A||B| + |A||\beta| + |B||\alpha|}{|b|^2}.$$

---

[1] See Chap. XIV, sec. xv.

But since

$$|b| \leqslant |B| + |\beta|,$$

and

$$|\beta| = |B - b| < \epsilon_2, \quad (\epsilon_2 \rightarrow 0)$$

we may replace $|b|^2$ in the right side of the above inequality by $|B|^2$. We have then

$$\frac{|a||b|}{|b|^2} < \frac{|A||B| + |A||\beta| + |B||\alpha|}{|B|^2},$$

$$\frac{|a|}{|b|} < \frac{|A|}{|B|} + \frac{|A||\beta| + |B||\alpha|}{|B|^2}.$$

But

$$\underset{\epsilon_3 \rightarrow 0}{\text{Limit}} |\alpha + \epsilon_3| = |\alpha_1|.$$

$$\underset{\epsilon_4 \rightarrow 0}{\text{Limit}} |\beta + \epsilon_4| = |\beta_1|.$$

Hence,

$$\underset{\substack{\epsilon_3 \rightarrow 0 \\ \epsilon_4 \rightarrow 0}}{\text{Limit}} \left| \frac{A(\beta + \epsilon_4) + B(\alpha + \epsilon_3)}{B^2} \right| \leqslant \frac{|A||\beta_1| + |B||\alpha_1|}{|B|^2},$$

which gives the relationship between the absolute error of the quotient of two approximate numbers a and b (when b is the divisor),

$$\frac{|A||\beta| + |B||\alpha|}{|B|^2},$$

and its upper limit.

The expression for the apparent error of the quotient of two approximate numbers is similar in form and content with the form and expression for the absolute error given above.

Since

$$0 < A_2 < A < A_1$$

and

$$0 < B_2 < B < B_1$$

we have

$$\frac{|A||\beta_1| + |B||\alpha_1|}{|B|^2} < \frac{|A_1||\beta_1| + |B_1||\alpha_1|}{|B_2|^2}.$$

A. THE RELATIVE ERROR

The relative error of two approximate numbers a and b (a the dividend, and b the divisor), whose respective relative errors are $\wp_a$ and $\wp_b$, is obtained as follows.

$$\frac{|\alpha|}{|A|} < \frac{|A||\beta| + |B||\alpha|}{|B|^2} : \frac{|A|}{|B|} = \frac{|\alpha|}{|A|} + \frac{|\beta|}{|B|} > \frac{|\beta|}{|B|},$$

or

$$|\wp_a| < |\wp_{(a:b)}| = |\wp_a| + |\wp_b| > |\wp_b|.$$

Similarly, we have

$$\wp'_a \leqslant \wp'_{(a:b)} \geqslant \wp'_b.$$

Finally, we obtain the relations

$$\wp'_{(a:b)} \leqslant \wp'_a + \wp'_b,$$

and

$$|\wp_{(a:b)}| \leqslant |\wp_a| + |\wp_b|.$$

From the above it follows that the quotient of two approximate numbers *a* and *b* (*a* the dividend, and *b* the divisor) should never be expected to be as accurate as the less accurate of the two approximate numbers. For example:

Approximate number	Apparent error	Relative error
642.6	+0.04	+0.000062
32.7	−0.02	−0.000612
19.65	+0.013	+0.000674

The resultant quotient is then 19.6 + 0.07 or 19.7 − 0.03.

B. THE DIRECTION OF THE ERRORS OF THE QUOTIENT OF TWO APPROXIMATE NUMBERS

The reader is referred to section IV, B of Chapter XVI, in which the case of a product of two approximate numbers whose absolute errors are of opposite directions is analyzed. If such a case arises in the process of division, the method of approach is the same.

We have that

$$0 < A_2 < A < A_1,$$
$$0 < B_2 < B < B_1,$$
$$a = A_1 - e = A_2 + e', \quad |e| \leqslant |e'| \quad \text{or} \quad |e'| \leqslant |e|,$$
$$b = B_1 - f = B_2 + f', \quad |f| \leqslant |f'| \quad \text{or} \quad |f'| \leqslant |f|.$$

When the directions of the apparent errors are the same, the quotient $(a:b)$ is

1. $$\frac{|a|}{|b|} < \frac{|A_1|}{|B_1|} + \frac{|A_1||f| + |B_1||e|}{|B_1|^2}, \quad \begin{matrix} |f| \leqslant |f'|, \\ |e| \leqslant |e'|, \end{matrix}$$

2. $$\frac{|a|}{|b|} < \frac{|A_2|}{|B_2|} + \frac{|A_2||f'| + |B_2||e'|}{|B_2|^2}, \quad \begin{matrix} |f'| \leqslant |f|, \\ |e'| \leqslant |e|. \end{matrix}$$

Replace B_1 by B_2 in 1 and A_2 by A_1 in 2 and their corresponding apparent errors. We have then

$$1. \qquad \frac{|a|}{|b|} < \frac{|A_1|}{|B_1|} + \frac{|A_1||f| + |B_1||e|}{|B_2|^2},$$

$$|f| \leqslant |f'|.$$

$$1.1. \qquad \frac{|a|}{|b|} < \frac{|A_1|}{|B_2|} + \frac{|A_1||f'| + |B_2||e|}{|B_2|^2}.$$

$$2. \qquad \frac{|a|}{|b|} < \frac{|A_2|}{|B_2|} + \frac{|A_2||f'| + |B_2||e'|}{|B_2|^2},$$

$$|e'| \leqslant |e|.$$

$$2.2. \qquad \frac{|a|}{|b|} < \frac{|A_1|}{|B_2|} + \frac{|A_1||f'| + |B_2||e|}{|B_2|^2}.$$

If we refer to the expression for the upper limit of the absolute error of the quotient of two approximate numbers (see page 171) we have that the upper limit of the absolute error, in its absolute value, is less than

$$\frac{|A_1||\beta_1| + |B_1||\alpha_1|}{|B_2|^2}.$$

The apparent error of the quotient $(a:b)$ in 1.1. and 2.2. has an absolute error such that it is

$$\leqslant \frac{|A_1||\beta_1| + |B_1||\alpha_1|}{|B_2|^2}.$$

The use of the upper limits of the true values of the approximate numbers for the dividend brings the apparent error of the quotient of the two approximate numbers closer to its upper limit.

It is important that we bear in mind the fact that a numerical representation of an approximate number possesses the properties of a rational representation of an irrational number (see page 107). The Dedekind cut of a difference of two irrational numbers is determined by two classes, namely:

$$\text{lower: } |A_2 - B_1|, \quad \text{and} \quad \text{upper: } |A_1 - B_2|.$$

The Dedekind cut of the quotient of two approximate numbers is determined by two classes:

$$\text{lower: } \frac{|A_2|}{|B_1|}, \quad \text{and} \quad \text{upper: } \frac{|A_1|}{|B_2|}.$$

Thus, it is obvious that in the case of subtraction and division the apparent errors of the approximate numbers should be opposite (before the processes take place).

Example:

Approximate number	Apparent error	Relative error
425.6	−0.03	−0.000070
341.7	−0.04	−0.000117
425.5	+0.07	+0.000164
341.7	−0.04	−0.000117
1.2452	+0.00035	+0.000281

The final result is then 1.246 − 0.0004.

C. THE ABSOLUTE AND RELATIVE ERRORS

Unlike the process of multiplication, the process of division does not offer a simple means for the determination of the magnitude of the apparent error, or absolute error, of the quotient of two approximate numbers.

The formula for the absolute (apparent) error of a quotient of two approximate numbers is very complicated from the computational point of view. Moreover, the division by $|B|^2$ is generally not a terminating process. Obviously, the use of the absolute (apparent) errors in the process of evaluating the approximativeness of a quotient of two approximate numbers is not economical either from the point of view of the time element or from the point of view of the labor element.

The nature and form of the relative error of the quotient of two approximate numbers, however, are such that a quick and direct method of analysis of the approximativeness of the quotient of the two approximate numbers may be developed. Moreover, by means of the formula[2]

$$z > y_1 10^k |\wp_{(a:b)}|$$

the apparent (absolute) error of the quotient may be determined.

1. *The relative error*

In the following analysis the approximate number a will always be the dividend, and the approximate number b will always be the divisor. Let the relative errors of a and b be $\wp_a$ and $\wp_b$ respectively.

We may have the following cases:

(1) a and b are given as correct to k significant digits each.

(2) a is given as correct to $(k + 1)$ significant digits, and b is given as correct to k significant digits.

(3) a is given as correct to k significant digits, and b is given as correct to $(k + 1)$ significant digits.

---

[2] See Chap. XIV, sec. xv.

(1) Let a and b be given as correct to k significant digits each. Here, we may have two cases:

(a)
$$\frac{\wp'_a}{\wp'_b} \eqslantless 10^{-r}, \quad (r = 0, 1),$$
that is,
$$|\wp_a| < |\wp_b|.$$

(b)
$$\frac{\wp'_b}{\wp'_a} \eqslantless 10^{-r}, \quad (r = 0, 1),$$
that is,
$$|\wp_a| > |\wp_b|.$$

There may be a third case, that is, when $|\wp_a| = |\wp_b|$. But the analysis of this case is similar to the above two, and we shall not examine it.

(a) Let $\quad |\wp_a| < |\wp_b|. \quad$ Then $\quad x_1 \geqslant y_1.$

We have then
$$|\wp_a| < \frac{1}{x_1 10^{k-1}}, \quad \text{and} \quad |\wp_b| < \frac{1}{y_1 10^{k-1}}.$$
Then
$$|\wp_{(a:b)}| \leqslant \frac{1}{x_1 10^{k-1}} + \frac{1}{y_1 10^{k-1}} = \frac{x_1 + y_1}{x_1 y_1 10^{k-1}}.$$
Also
$$\frac{x_1 + y_1}{x_1 y_1 10^{k-1}} > \frac{1}{y_1 10^{k-1}}.$$
Also
$$\frac{x_1}{x_1 + y_1} \leqslant w10^{-1}, \quad (w = 1, 2, 3, \cdots, 9).$$
Hence,
$$|\wp_{(a:b)}| \leqslant \frac{1}{wy_1 10^{k-2}}.$$

This shows that the relative error $\wp_{(a:b)}$ of the quotient is of the order either "zero" or "one" in comparison with the relative error of the less accurate number b, all depending on whether $wy_1 \geqslant 10$, or $wy_1 < 10$.

(b) If
$$|\wp_a| > |\wp_b|, \quad \text{then} \quad x_1 \leqslant y_1, \quad \text{and}$$
$$\frac{x_1 + y_1}{x_1 y_1 10^{k-1}} > \frac{1}{x_1 10^{k-1}}.$$
Then
$$\frac{y_1}{x_1 + y_1} \leqslant w10^{-1}, \quad (w = 1, 2, 3, \cdots, 9).$$

Hence,

$$|\wp_{(a:b)}| \leqslant \frac{1}{wx_1 10^{k-2}},$$

and the relative error $\wp_{(a:b)}$ of the quotient is of the order either "zero" or "one," in comparison with the relative error of the less accurate number a, all depending on whether $wx_1 \geqslant 10$, or $wx_1 < 10$.

The reader will find no difficulty in verifying the above results by considering the specific values of x_1 and y_1 and the respective apparent errors of the approximate numbers.

(2) Let a be given as correct to $(k + 1)$ significant digits, and b be given as correct to k significant digits. Then $|\wp_a| \leqslant |\wp_b|$, and $x_1 \gtreqless y_1$. We have then

$$|\wp_a| < \frac{1}{x_1 10^k}, \quad \text{and} \quad |\wp_b| < \frac{1}{y_1 10^{k-1}}.$$

Then

$$|\wp_{(a:b)}| \leqslant \frac{1}{x_1 10^k} + \frac{1}{y_1 10^{k-1}} = \frac{10x_1 + y_1}{x_1 y_1 10^k}.$$

But

$$\frac{10x_1 + y_1}{x_1 y_1 10^k} > \frac{1}{y_1 10^{k-1}}.$$

Also

$$\frac{x_1 y_1}{10x_1 + y_1} \leqslant w10^{-1}, \quad (w = 1, 2, 3, \cdot \cdot \cdot, 9).$$

Hence,

$$|\wp_{(a:b)}| \leqslant \frac{1}{w10^{k-1}},$$

and the relative error of the quotient of the two approximate numbers is of the order "zero" in comparison with the relative error of the less accurate approximate number b.

The reader will find no difficulty in verifying the above result by considering the specific values of x_1 and y_1 and the apparent errors of the respective approximate numbers. But he should bear in mind that in this case

$$\frac{\wp'_a}{\wp'_b} \approxeq 10^{-r}, \quad (r = 0, 1, 2).$$

(3) Let a be given as correct to k significant digits, and b be given as correct to $(k + 1)$ significant digits.

In this case

$$\frac{\wp'_b}{\wp'_a} \approxeq 10^{-r}, \quad (r = 0, 1, 2).$$

Also

$$|\wp_a| > |\wp_b|, \quad \text{and} \quad x_1 \gtreqless y_1.$$

Then

$$|\wp_a| < \frac{1}{x_1 10^{k-1}}, \quad \text{and} \quad |\wp_b| < \frac{1}{y_1 10^k}.$$

Then we have

$$|\wp_{(a:b)}| \leqslant \frac{1}{x_1 10^{k-1}} + \frac{1}{y_1 10^k} = \frac{x_1 + 10y_1}{x_1 y_1 10^k}.$$

But

$$\frac{x_1 + 10y_1}{x_1 y_1 10^k} > \frac{1}{x_1 10^{k-1}}.$$

Also

$$\frac{x_1 y_1}{x_1 + 10y_1} \leqslant w10^{-1}, \quad (w = 1, 2, 3, \cdots, 9).$$

Hence

$$|\wp_{(a:b)}| \leqslant \frac{1}{w10^{k-1}},$$

and this shows that the relative error $\wp_{(a:b)}$ is of the order "zero" in comparison with the relative error of the less accurate approximate number a.

The reader will find no difficulty in verifying the above result by considering the specific values of x_1 and y_1 and the respective apparent errors of the approximate numbers.

D. SIGNIFICANT DIGITS

Unlike the process of multiplication, in the process of division it is absolutely essential to know in advance "how far the division should be carried." This is because division is usually a non-terminating process. The information necessary in this case should relate to the place of significance in the quotient where the first significant digit on the left of the absolute (apparent) error appears. This information is secured by means of the formula[1]

$$z > y_1 10^k |\wp_{(a:b)}|.$$

The results of section II, C, *i* show that the quotient of two approximate numbers a and b (a the dividend, and b the divisor) which are correct to k significant digits each, is correct to k or $(k - 1)$ significant digits.

---

[1] See Chap. XIV, sec. XV.

Particular attention is necessary when the apparent errors of the approximate numbers are not given, or when the first significant digit of both these apparent errors is 5. This case was thoroughly examined in Chapter XVI, pages 160 to 162, for multiplication. Since the relative error of a quotient of two approximate numbers is of the same form as the relative error of a product of two approximate numbers, it is safe to assume that whatever results have been obtained in the case of multiplication will hold good for the case of division. We then can formulate the following.

If two approximate numbers are both given as correct to k significant digits, and the apparent errors for both are such that the first significant digit on the left is 5 (or if the apparent errors are not given, then it is assumed that the respective apparent errors are ± 0.5 of a unit to which the respective approximate numbers are correct), then if the product of the first significant digits on the extreme left of each approximate number is ≥ 10, the quotient is correct to k significant digits. Otherwise, the quotient is correct to $(k - 1)$ significant digits. If there is doubt about the product of the digits, then the product of the first two significant digits on the extreme left of the two approximate numbers should be examined to determine whether it is $\geq 1{,}000$ (then the quotient is correct to k significant digits); otherwise, the quotient is correct to $(k - 1)$ significant digits.

The same criterion holds for the case when the two approximate numbers are given with the type of apparent error described above, and one number is correct to $(k + 1)$ significant digits, whereas the other is correct to k significant digits. *It is immaterial which of the approximate numbers is the dividend and which is the divisor.*

The advisability of keeping, if possible, an extra significant digit in one of the approximate numbers whose quotient is to be obtained, may be seen if we examine the relative errors of the quotients when

(1) one of the approximate numbers is given as correct to $(k + 1)$ significant digits, and the other is given as correct to k significant digits, that is,

$$|\varphi_{quot}| \leq \frac{10x_1 + y_1}{x_1 y_1 10^k}, \quad \text{or} \quad |\varphi_{quot}| \leq \frac{x_1 + 10y_1}{x_1 y_1 10^k};$$

and

(2) the two approximate numbers are both given as correct to k significant digits.

The reader will find no difficulty in verifying the fact that the relative error of a quotient of two approximate numbers in the first case (1) is smaller than the relative error of the quotient of the same two approximate numbers when the more accurate one is rounded off to k significant digits as in the second case (2).

E. THE RULE FOR DIVISION

Division of approximate numbers should be performed in accordance with the following rules.

1. If the two approximate numbers whose quotient is to be obtained do not both have the same number of significant digits, then the more accurate approximate number should be rounded off so that it will have one more significant digit than the less accurate approximate number.

2. If the quotient of two approximate numbers is known to be correct to k significant digits, then the process of division should be carried to $(k + 1)$ digits.

3. The relative error of the quotient of two approximate numbers may be equal to the sum of the relative errors of the respective approximate numbers and the directions of these errors must be the same.

4. If the directions of the relative (and apparent) errors of the approximate numbers whose quotient is to be obtained, are the same, then the limit of the true value of one approximate number should be changed to the upper limit or the lower limit, and this will make the directions of both errors (relative as well as apparent) opposite. The direction of the relative error of the quotient of two approximate numbers is the same then as the direction of the respective apparent errors of the dividend.

5. If one of the approximate numbers has a positive (or negative) error, and the other approximate number has an error with the sign ±, then the process of division should be carried out by the method of bounds.[4]

6. If two approximate numbers that are to be divided are such that the less accurate contains k significant digits, and no information concerning the apparent and (or) the relative errors of the approximate numbers is given, then the resultant quotient of the two approximate numbers will contain k significant digits if and only if the product of the first significant digits on the extreme left of each of the approximate numbers is ≥ 10, or if in doubt, the product of

---

[4] See Chap. XI.

DOCUMENT RESUME

ED 096 173 SE 018 473

AUTHOR	Bakst, Aaron
TITLE	Approximate Computation. National Council of Teachers of Mathematics, Yearbook 12 [1937]. Reprint 1966.
INSTITUTION	National Council of Teachers of Mathematics, Inc., Washington, D.C.
PUB DATE	[37]
NOTE	296p.; Missing pages in bibliography section
AVAILABLE FROM	National Council of Teachers of Mathematics, Inc., 1906 Association Drive, Reston, Virginia 22091
EDRS PRICE	MF-$0.75 HC Not Available from EDRS. PLUS POSTAGE
DESCRIPTORS	Algebra; *Curriculum; Guides; *Instruction; Mathematical Concepts; *Mathematics Education; Number Concepts; Reference Books; *Secondary School Mathematics; Trigonometry; *Yearbooks
IDENTIFIERS	*Computation; Estimation; Numerical Methods

ABSTRACT

Divided into two major parts, the teaching of approximate computation and the theory of numerical processes, this yearbook has four overall goals. The first is to determine to what extent the recommendations concerning approximate computation, which were made by the National Committee on Mathematical Requirements, have been carried out. The second is the analysis of the theory of numerical processes with suitable mathematical rigor. The third is the development of proper techniques for working with approximate data in the secondary schools. The last is the presentation of outlines guiding the treatment of approximate computation in elementary algebra and trigonometry, intermediate algebra, and advanced algebra and trigonometry. (LS)

ED 096173

THE NATIONAL COUNCIL OF TEACHERS OF MATHEMATICS

THE TWELFTH YEARBOOK

APPROXIMATE COMPUTATION

By Aaron Bakst, Ph.D.

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL BY MICROFICHE ONLY HAS BEEN GRANTED BY
Charles R. Hucka
N. C. T. M.
TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER

SE 018 473

AMS REPRINT COMPANY
NEW YORK

Copyright by
THE NATIONAL COUNCIL OF TEACHERS
OF MATHEMATICS

Reprinted with the permission of
the original publisher

AMS REPRINT COMPANY
A Trade Division of
AMS PRESS, INC.
NEW YORK, N.Y. 10003
1966

Manufactured in the United States of America

EDITOR'S PREFACE

THIS is the twelfth of a series of Yearbooks which the National Council of Teachers of Mathematics began to publish in 1926. The titles of the first eleven Yearbooks are as follows:

1. A Survey of Progress in the Past Twenty-Five Years.
2. Curriculum Problems in Teaching Mathematics.
3. Selected Topics in the Teaching of Mathematics.
4. Significant Changes and Trends in the Teaching of Mathematics Throughout the World Since 1910.
5. The Teaching of Geometry.
6. Mathematics in Modern Life.
7. The Teaching of Algebra.
8. The Teaching of Mathematics in the Secondary School.
9. Relational and Functional Thinking in Mathematics.
10. The Teaching of Arithmetic.
11. The Place of Mathematics in Modern Education.

Bound copies of all except the first two Yearbooks can be secured from the Bureau of Publications, Teachers College, Columbia University, New York, N. Y., for $1.75 each postpaid. The first Yearbook is now out of print and the second is obtainable only in paper covers ($1.25 postpaid). A complete set of Yearbooks (Numbers 2 to 12 inclusive) will be sent postpaid for $15.

Since the tendency of some of the newer textbooks in mathematics for the secondary school is to introduce the method of Approximate Computation at one place or another, it has seemed desirable to present a fairly comprehensive treatment of the method so that teachers who may wish to get help will find this Yearbook useful.

I wish to express my personal appreciation as well as that of the National Council of Teachers of Mathematics both to Dr. Aaron Bakst for permitting us to publish his contribution as the Twelfth Yearbook and to Professor A. A. Bennett for his Foreword and the great interest that he has shown in the publication of this volume.

W. D. REEVE

THE NATIONAL COUNCIL OF TEACHERS OF MATHEMATICS

OFFICERS

Honorary President—Herbert E. Slaught, University of Chicago, Chicago.

President—Martha Hildebrandt, Proviso Township High School, Maywood, Ill.

First Vice-President—Mary Kelly, Wichita, Kans.

Second Vice-President—J. T. Johnson, Chicago, Ill.

Secretary-Treasurer—Edwin W. Schreiber, Western Illinois State Teachers College, Macomb, Ill.

EDITORIAL COMMITTEE

Editor-in-Chief and Business Manager of The Mathematics Teacher and the Yearbook—William David Reeve, Teachers College, Columbia University, New York City.

Associate Editors—Vera Sanford, State Normal School, Oneonta, N. Y.; W. S. Schlauch, Hasbrouck Heights, N. J.

ADDITIONAL MEMBERS ON THE BOARD OF DIRECTORS

One Year	M. L. Hartung, Madison, Wis.	1938
	Mary A. Potter, Racine, Wis.	1938
	Rolland R. Smith, Springfield, Mass.	1938
Two Years	E. R. Breslich, Chicago, Ill.	1939
	Virgil Mallory, Montclair, N. J.	1939
	Leonard D. Haertter, Clayton, Mo.	1939
Three Years	Wm. Betz, Rochester, N. Y.	1940
	H. C. Christofferson, Oxford, Ohio	1940
	Edith Woolsey, Minneapolis, Minn.	1940

The National Council of Teachers of Mathematics is a national organization of mathematics teachers whose purpose is to:

1. Create and maintain interest in the teaching of mathematics.
2. Keep the values of mathematics before the educational world.
3. Help the inexperienced teacher to become a good teacher.
4. Help teachers in service to become better teachers.
5. Raise the general level of instruction in mathematics.

Anyone interested in mathematics is eligible to membership in the National Council upon payment of the annual dues of $2.00. The dues include subscription to the official journal of the Council (*The Mathematics Teacher*). Correspondence should be addressed to *The Mathematics Teacher*, 525 West 120th Street, New York City.

FOREWORD

The exactness of mathematics has been considered one of its conspicuous virtues. The certainty of the results of elementary mathematical computation sets this subject apart from any of the following: (1) the natural sciences with their empirical estimates and inductive generalizations; (2) jurisprudence and theology with their bases in authority, human and divine; (3) economic, meteorological, or political forecasting with their dependence upon personal judgment; (4) history with its fabric interwoven with the strand of time. Mathematics is universal as to time, place, race, and application. Thus when computation far outstrips possible direct measurement, as when the decimal expansion of *Pi* is correctly computed to hundreds of significant figures, one must retain confidence in this achievement of the human spirit or else doubt the validity of all rational processes.

Although a teacher may sometimes slight the spiritual claims of mathematical development while laboring to impress the pupils' minds with the essentials of operational technique, he cannot, even in the midst of drill and the routine of explanation, wholly lose sight of the logical certainty inherent in correct deduction. However, as experience has shown, the subject of *approximate computation* may remain wholly unmentioned, and the pupil, who as he grows older will find the use of approximations increasingly necessary, may be equipped with only a rudimentary notion of the mathematical relations among measured quantities.

Quantitative data, whether arising from (1) physical measurement, (2) estimates, (3) condensation of more elaborate lists, (4) tables rounded off at a given decimal place, or otherwise, may be regarded as "correct" and yet enjoy precision of a necessarily limited sort. To emphasize a distinction not always clearly stated: numbers may be *precise* to within intervals of a given size, such as to the nearest mile, the nearest tenth of a second, or the nearest ten-thousandth of a unit; or again, numbers may be *accurate* to a given per cent or to a given number of significant digits—although this special use of words is not universally adopted. For addition and subtraction, numbers are furnished preferably with a given precision, since further precision for a single term in a sum remains at best ineffective. On the other hand, a comparable relative accuracy is to be desired for factors in a product. It seems not inappropriate to record the necessarily approximate results of experimental measurement as listed, with an estimate of the maximum error. In this way a reported value of, say, 235.72 might be correct to the nearest hundredth of a unit and thus lie between 235.715 and 235.725, or, say, to the nearest hundredth of one per cent and thus lie between 235.72 (99.99%) = 235.70 −, and 235.72 (100.01%) = 235.74 +. In spite of the fact that teachers have long admitted the importance of a clear grasp of these applications of mathematical common sense, it seems unquestionable that practice in elementary school instruction has not always borne evidence of this appreciation. Although perhaps little time should be devoted to approximate numbers as an isolated classroom topic, the whole course in elementary mathematical study after about the sixth year should *reflect* a readiness to accept results of observed measurement at their true and limited worth. The teacher may well spend considerable time at some stage of his study in straightening out his own ideas and in acquiring a justified confidence without which clarity of exposition is not to be expected. It is to this end that this Yearbook is presented.

Perhaps a special remark is in order with regard to such a notation as 235.72 ± 0.005. This might conceivably be employed to indicate the mean value of several observations together with the total range within which deviation might be viewed as having occurred. Such a notation might seem appropriate when discussing forcing errors and rounding off numbers in a tabulated list. In practice, however, throughout most of the applications to natural science a different convention is adopted. There the correction number appended after the sign, ±, is not the semi-range but the approximate computed "probable error." The rules for operation with probable errors are based upon an assumed approximate applicability of the normal law of distribution of deviations, and are of course notably different from those with which this Yearbook is concerned.

When exploring the extensive discussion of numerical precision and accuracy presented by Dr. Bakst, the reader should not forget the setting of this study among other aspects of the general problem of the mathematics of approximation. These are the finite ("one-step") approximations such as $22/7$ for π, $1 + (a/2)$ for $\sqrt{(1 + a)}$, Simpson's rule for areas, and the result of linear interpolation for the value of a function between listed values of the argument. In contrast to these there are also formulas and algorithms for successively improved approximations, invoking eventually the entire theory of infinite numerical processes. But methods of approximation in their turn treat only part of the topic. One faces also the problem of the use and interpretation of approximate data, as developed in the theory of statistics—a subject whose manifold aspects and practical contacts need no word of appreciation here.

While we all admire the purity of our fair science, let us not be dazzled by its white brilliance, but learn to appreciate its facile flexibility, its adaptability to whatever may be the special purpose in mind. Mathematics does not always dictate what shall be the form of the answer. Let us learn to develop common sense in stating and interpreting problems and to apply the appropriate mathematical tools in finding a quantitative answer to suitable and reasonable approximativeness. Let not mathematical technique smother common sense; may its study develop not pedantry but power.

Albert A. Bennett

Brown University

PREFACE

It is the aim of the writer to present in this work a rigorous treatment of the methods of computation with approximate numbers. Naturally, before any approach from the mathematical point of view can be made, it is necessary to define the elements that are dealt with. Two types of quantities find place in mathematics rather early in the elementary stages of the subject. These two types always go together, constantly entering more vividly into the experience of the student of mathematics. One type is abstract, physically non-existent, elusive, but steadily growing in importance in every branch of mathematics. The other is real, a product of practical experience, a thing "one can lay one's hands on," so to speak. One type is existent as an abstraction only, and the other represents quantities that tend to represent these abstractions. One type embraces everything that is exact, and the other contains the prototypes of these exact quantities.

In practical mathematics, that is in that branch of mathematics in which numerical processes with quantities of these two types arise, the methods of obtaining satisfactory numerical results from the operations with these quantities must be considered. "Satisfactory results" may not mean satisfactory to the practical mathematician from his individual point of view. The personal element should never find a place in any mathematical process. Satisfactory results imply that there must be some means and methods by which numerical processes can be applied to quantities of the two types referred to, and results obtained consistent with the nature of the quantities.

In the actual experience of the practical mathematician the quantities of the second type are the predominant ones. Even though "exact" quantities are always present, the actual numerical process takes place with another quantity which *approximates* it to some extent.

Thus we have two classes of quantities: exact quantities and approximate quantities. In actual experience the approximate quantity usually comes first. It is only in the stage of abstraction that the exact quantity, the approximation of which is known, is considered. This method of development is followed in this work in the process of the analysis of the relationship between exact and approximate numbers.

In order to present a purely mathematical treatment of the subject of Approximate Computation, approximate numbers are considered in the most general form. This fact necessitated the introduction of "exact values" and "absolute errors" whose prototypes in practical numerical processes are "true values" and "apparent errors." The passing from the general formula to "the practical for computation purposes" formula, therefore, requires no special explanations.

The "absolute error" should be thought of as an abstraction of the "apparent error" at a certain stage of the process of approximation. The "upper limit of the absolute error" is akin in its nature to the so-called upper limit of the upper bound of a function. This term is of theoretical importance in the processes of the analysis of Approximate Computation.

There are three distinct approaches to the problem of Approximate Computation. The basic, classical approach represents the analysis of the errors of approximate numbers and their influence on the results of the performed numerical processes. Results of numerical operations usually contain three types of digits, namely: significant digits, those that are definitely worthless, and those (in intermediate position) that are semi-reliable, so to speak. The third type of digits, that is, the semi-reliable ones may be at some time of value to the computer. This study, however, concerns itself only with the determination of the first type of digits, that is, significant digits. The fundamental problem of this method of analysis is the evaluation of the maximum error in each particular case. This method of analysis is pursued in the present volume. Of the works on this phase of the subject the most widely known are by Fassbinder, Lüroth, and Kavun.

At times the evaluation of the bounds between which an approximate number is located is preferred to the evaluation of the errors. Both the method of errors and the method of bounds are favored by computers of tables as well as by all who are interested in numerical results strictly requiring correctness.

The second approach to the problem of Approximate Computation dispenses entirely with the evaluation of the errors. This method requires the writing of an approximate number with one extra digit to the right of the last significant digit, and this extra digit actually corresponds to the apparent error. Theoretically this phase of Approximate Computation is based on the third method of analysis described below. Of the works along these lines of analysis the most thorough are by Bourget, Kryloff, Van der Flit, and Selivanoff.

The third approach to the problem of Approximate Computation is based on the Theory of Probability. This phase of analysis is concerned not alone with the evaluation of errors; the primary concern is centered on the question of the probability of the occurrences of these errors. Theoretically, study of this phase of analysis is still not completed. Much remains to be done in this field from the point of view of pure mathematical research. Of the works along these lines can be mentioned those of Bradis, Stadthagen, and Volberg.*

Some of the topics discussed in this volume have never before been treated from the point of view of Approximate Computation. It is the aim of the writer to cover the field of elementary mathematics in such a way as to arouse the interest in the subject of Approximate Computation which rightfully belongs to it.

This preface would not be complete without the expression of the writer's grateful thanks to Dr. William D. Reeve, Professor of Mathematics, Teachers College, Columbia University and to Dr. Edward Kasner, Professor of Mathematics, Columbia University, without whose inspiring encouragement and help this study would have neither been attempted nor completed. Professors Clifford B. Upton and Helen M. Walker, both of Teachers College, Columbia University, and Professor Albert A. Bennett of Brown University read the manuscript and offered many valuable and critical suggestions. Professors John R. Clark and William R. Odell, both of Teachers College, Columbia University, read the manuscript and offered many valuable criticisms and suggestions to Part I. To Dr. John A. Swenson, Associate in Mathematics, Teachers College Columbia University, and Professor Carl N. Shuster, State Teachers College, Trenton, New Jersey, the writer is indebted for many practical suggestions. The enumeration of many others of the writer's friends and colleagues who evinced genuine interest in this study would require too much space, and therefore he offers his thanks to all of those without naming them. To Miss Eloise L. Bernstein the writer is indebted for her help in the final preparation of the manuscript for publication. And last, but not least, the writer is indebted to his mother for her encouragement and inspiration while this study was in progress.

A. B.

---

* A bibliography on Approximate Computation is given in Appendix 4.

CONTENTS

PART I

THE TEACHING OF APPROXIMATE COMPUTATION

CHAPTER PAGE

I. INTRODUCTION TO PART I 3

I. Approximate Computation in Elementary Mathematics Curriculum. II. The Analysis of American Textbooks. III. The Status of Approximate Computation in Foreign Countries.

II. APPROXIMATE COMPUTATION IN SECONDARY SCHOOL MATHEMATICS 13

III. DEVELOPMENT OF APPROXIMATE COMPUTATION IN SECONDARY SCHOOL MATHEMATICS 17

IV. APPROXIMATE COMPUTATION IN ELEMENTARY ALGEBRA AND NUMERICAL TRIGONOMETRY: MEASUREMENT 19

I. The Development of the Notion of Measurement. II. The Process of Measurement.

V. APPROXIMATE COMPUTATION: APPROXIMATE NUMBERS 31

I. The Origin and Kinds of Approximate Numbers. II. The Error in Measurement as a Part of an Approximate Number. III. Significant Digits. IV. Rounding Off Digits. V. The Relative Error. VI. The Relation Between the Magnitudes of Approximate Numbers and the Magnitudes of Their Respective Apparent Errors.

VI. APPROXIMATE COMPUTATION: NUMERICAL OPERATIONS 47

I. Addition. II. Subtraction. III. Multiplication. IV. Division. V. Raising to a Power. VI. Extraction of Roots.

VII. APPROXIMATE COMPUTATION: LOGARITHMS OF NATURAL NUMBERS 72

I. The Approximate Nature of the Numerical Values of Logarithms of Natural Numbers and Especially of Approximate Numbers. II. The Relation of the Approximativeness of Logarithms and Their Corresponding Approximate Numbers. III. Tables of Logarithms.—Interpolation.

VIII. APPROXIMATE COMPUTATION: NATURAL TRIGONOMETRIC FUNCTIONS 75

I. The Approximate Nature of the Numerical Values of

CHAPTER PAGE

Natural Trigonometric Functions. II. The Relation Between the Approximativeness of an Angle and Its Corresponding Natural Trigonometric Functions. III. The Use of Numerical Values of Natural Trigonometric Functions in Connection with Other Approximate Numbers. IV. Tables of Natural Trigonometric Functions.—Interpolation. V. Determination of an Approximate Angle Whose Natural Trigonometric Function Is Given.

IX. APPROXIMATE COMPUTATION. APPROXIMATE COMPUTATION OF THE FIRST TYPE 79

I. The Notion of Optimum Approximativeness. II. Formulas. III. Equations.

X. APPROXIMATE COMPUTATION: APPROXIMATE COMPUTATION OF THE SECOND TYPE 84

I. The Notion of Predetermined Approximativeness. II. Determination of Approximativeness of Data in Order to Comply with Limitations of Approximativeness of the Result Set Forth in the Problem. III. Determination of Approximativeness of Measurements to Be Performed in Order to Obtain a Final Numerical Result of Predetermined Approximativeness.

XI. APPROXIMATE COMPUTATION IN INTERMEDIATE ALEGBRA 87

I. General Remarks. II. Representation of Approximate Numbers. III. Rounding Off Digits. IV. Numerical Processes. V. Evaluation of Formulas.

XII. APPROXIMATE COMPUTATION IN ADVANCED ALGEBRA AND TRIGONOMETRY 95

I. General Remarks. II. Numerical Processes. III. Logarithmic Functions.

PART II

THEORY OF NUMERICAL PROCESSES

XIII. INTRODUCTION TO PART II 105

The Nature of Approximate Computation.

XIV. APPROXIMATIONS 112

I. Exact and Approximate Quantities. II. The Approximativeness. III. Formulas. IV. Criteria for Determination of Approximativeness. V. Numerical Data. VI. Approximate Numerical Data. VII. The Criterion for Approximations. VIII. The Decimal Place Approximativeness. IX. The Significant Digits Approximativeness. X. The Approximativeness of Numerical Results. XI. The Apparent Error. XII. The Exact Value of an Approximate Number and the Absolute Error. XIII. The Relative Error. XIV. Precision

CHAPTER PAGE

and Accuracy. XV. Absolute (Apparent) and Relative Errors. XVI. The Order of Errors. XVII. The Representation of Approximate Numbers.

XV. THE METHODS OF APPROXIMATE COMPUTATION 133

I. The Two Types of Problems in Approximate Computation. II. The Approximativeness of Numerical Results.

XVI. APPROXIMATE COMPUTATION OF THE FIRST TYPE 137

I. General Remarks. II. Addition. III. Subtraction. IV. Multiplication.

XVII. APPROXIMATE COMPUTATION OF THE FIRST TYPE (Continued) 165

I. Raising to a Power. II. Division. III. Extraction of Roots.

XVIII. APPROXIMATE COMPUTATION OF THE FIRST TYPE (Concluded) 190

I. Approximate Computation with Exact Numbers. II. Complex Approximate Computation. III. The Approximativeness of Roots of Equations. IV. Approximate Computation with Approximate Numbers Which Are Given as Correct Within a Relative Error. V. Approximate Computation with Approximate Numbers Which Are Given as Correct Within a Stated Number of Significant Digits. VI. Summary.

XIX. APPROXIMATE COMPUTATION OF THE SECOND TYPE .. 216

I. The Problem. II. Addition. III. Subtraction. IV. Multiplication. V. Division. VI. Raising to a Power. VII. Extraction of Roots. VIII. Complex Approximate Computation. IX. The Determination of Data Whose Approximativeness Is Given in Advance.

XX. LOGARITHMS 229

I. The Method of Proportional Parts. II. The Precision of Tables of Logarithms. III. Operations with Logarithms. IV. The Absolute Error of the Logarithm of an Approximate Number.

XXI. TRIGONOMETRIC FUNCTIONS 243

I. The Method of Proportional Parts. II. Accuracy of Tables of Natural Trigonometric Functions. III. The Absolute Error of a Trigonometric Function.

XXII. HISTORY OF APPROXIMATE COMPUTATION 260

I. General Remarks. II. The Notation for Approximate Numbers. III. The "Essai sur Les Nombres Approximatifs." IV. Gauss's Theoria Motus. V. The Law of Errors. VI. The Introduction of the Relative Error in the Processes of Approxi-

CHAPTER PAGE
mate Computation. VII. The Analysis of Contracted Multiplication and Division.

APPENDIX

1. Rounding Off Digits 275
I. General Remarks. II. The Nature of an Approximate Number. III. Rounding Off Approximate Numbers. IV. The Relative Error. V. Rounding Off Approximate Numbers Whose Errors Have the Sign ±.

2. Note on Antilogarithms 280

3. Symbols Used in the Text 282

4. Bibliography 284

PART I

THE TEACHING OF APPROXIMATE COMPUTATION

1/2

CHAPTER I

INTRODUCTION TO PART I

1. APPROXIMATE COMPUTATION IN ELEMENTARY MATHEMATICS CURRICULUM

IN its Report of 1923, the National Committee on Mathematical Requirements recommended the inclusion of topics on Approximations and Approximate Computation. The committee clearly indicated the importance of this phase of mathematical instruction, and made general suggestions concerning it.

> . . . We may now approach the problem of formulating the more important aims that the teaching of mathematics should serve. It has been customary to distinguish three classes of aims: (1) practical or utilitarian, (2) disciplinary, (3) cultural; and such a classification is indeed a convenient one. It should be kept clearly in mind, however, that the three classes mentioned are not mutually exclusive and that convenience of discussion rather than logical necessity often assigns a given aim to one or the other of these classes. Indeed, any truly disciplinary aim is practical and, in a broad sense, the same is true of cultural aims.[1]

Elaborating on the practical aims, the Report continues:

> (*b*) Exercise of common sense and judgment in computing from approximate data, familiarity with the effect of small errors in measurement, the determination of the number of figures to be used in computing and to be retained in the result, and the like.[2]

In the outlines of subject matter for grades seven, eight, and nine (the junior high school) the Report, under the subtitle "Intuitive Geometry," recommends the teaching of:

> (*a*) The direct measurement of distances and angles by means of a linear scale and protractor. The approximate character of measurement. An understanding of what is meant by the degree of precision as expressed by the number of "significant" figures.
>
> (*c*) Practice in numerical computation with due regard to the number of figures used or retained.[3]

---

[1] The Reorganization of Mathematics in Secondary Education, p. 6. A Report by the National Committee on Mathematical Requirements, The Mathematical Association of America, Inc., 1923.

[2] *Ibid.*, p. 7.

[3] *Ibid.*, p. 22.

Under the subtitle "Numerical Computation, Use of Tables, etc.," the Report states:

> The solution of problems should offer opportunity throughout the grades under consideration for considerable arithmetical and computational work. In this connection attention should be called to the importance of exercising common sense and judgment in the use of approximate data, keeping in mind the fact that all data secured from measurement are approximate. A pupil should be led to see the absurdity of giving the area of a circle to a thousandth of a square inch when the radius has been measured only to the nearest inch. He should understand the conception of "the number of significant figures" and should not retain more figures in his result than are warranted by the accuracy of his data. The ideals of accuracy and of self-reliance and the necessity of checking all numerical results should be emphasized. An insight into the nature of tables, including some elementary notions of interpolation, is highly desirable. The use of tables of various kinds (such as squares and square roots, interest, and trigonometric functions) to facilitate computation and to develop the idea of dependence should be encouraged.[4]

In another place[5] the Report mentions the fact that in the findings of a questionnaire study, computation with approximate data—rational use of significant figures—was regarded by 97 per cent of the persons responding as essential and of considerable value as a preparation for elementary college courses. A detailed analysis of this questionnaire study[6] reveals the fact that 97 per cent of the persons responding considered this topic essential and of considerable value as a preparation for college courses in physical sciences; 67 per cent considered it essential and of considerable value, and 20 per cent considered it as of some value, as a preparation for college courses in social sciences. Since then the College Entrance Examination Board has required of the candidates that they shall be able "to use four-place tables of natural trigonometric functions; that the pupils should acquire facility in simple interpolation; in general, emphasis should be laid on carrying the computation to the limit of accuracy permitted by the table."[7]

The inclusion of Approximate Computation in senior high school mathematics is recommended by the Report as follows:

> The solution of problems should offer opportunity throughout the course for considerable arithmetical and computational work. The conception of

---

[4] *Ibid.*, p. 28. [5] *Ibid.*, p. 46. [6] *Ibid.*, p. 45.

[7] College Entrance Examination Board: Definition of the Requirements in Elementary Algebra, Advanced Algebra, and Trigonometry, p. 4. Document No. 107, 1929.

> algebra as an extension of arithmetic should be made significant both in numerical applications and in elucidating algebraic principles. Emphasis should be placed upon the use of common sense and judgment in computing from approximate data, especially with regard to the number of figures retained, and on the necessity for checking the results. The use of tables to facilitate computation (such as tables of squares and square roots, of interest, and of trigonometric functions) should be encouraged.[5]

These recommendations are sufficiently emphatic to influence textbook writers and teachers of mathematics to give due place to Approximate Computation and Approximations in secondary school mathematics. In view of the modern spread of the applications of mathematics to all fields of human endeavor, the spirit in which Approximations and Approximate Computation have been recommended by the National Committee on Mathematical Requirements should give these topics a rather prominent place in the curriculum. The recommendations of the National Committee should by no means lead one to lay stress only on the mechanical phases of Approximate Computation. Ideally, Approximations and Approximate Computation should be interwoven whenever and wherever feasible with the entire content of secondary school mathematics.

In general, the Report of the National Committee has been adopted throughout this country as a guiding instrument in the construction of courses in mathematics and for writers of mathematics textbooks. The recommendations of this committee relating to Approximations and Approximate Computation, although general in scope, are of such importance that one would expect that due regard and attention would be given to them by both textbook writers and teachers of mathematics. It is the purpose of this study to determine the status of the teaching of Approximations and Approximate Computation in this country and to develop means by which the above recommendations can be carried out.

The scope of this study is fourfold, including:

1. The determination of the extent to which the recommendations of the National Committee, as far as Approximations and Approximate Computation are concerned, have been carried out.
2. The analysis and development of the Theory of Numerical Processes from the point of view of mathematical rigor.
3. The development of the proper mathematical technique for working with approximate data suitable for secondary school purposes.

---

[5] National Committee on Mathematical Requirements: *op. cit.*, p. 50.

4. The presentation on the basis of the results secured in procedures 2 and 3 above of outlines for the development and presentation of the subject of Approximations and Approximate Computation for: (*a*) elementary algebra and numerical trigonometry; (*b*) intermediate algebra; and (*c*) advanced algebra and trigonometry.

Part I of this study gives the results of an examination of the content of textbooks in elementary mathematics published in this country between 1923 and 1932. This phase of the investigation covers not only the direct study of the content of textbooks, but also the textual material and treatment of approximate numbers, even though no direct mention of the approximate nature of the data is made in the subject matter presented.

The proper teaching of any subject is possible only when the teacher is thoroughly familiar with the subject and has a genuine command of it. The subject of Approximations and Approximate Computation, especially within the scope of the Theory of Numerical Processes, has never been fully developed, nor has it ever been presented in a unified form. The teacher training and other institutions of higher education in this country have never provided for the academic training and the professional preparation of teachers of mathematics in the field of Approximations and Approximate Computation. Part II of this investigation is devoted to the development of the Theory of Numerical Processes which underlies the theory of Approximate Computation. The Theory of Numerical Processes is developed from the elementary point of view, but with a rigorous mathematical treatment throughout, suitable for the academic and professional training of teachers of mathematics in the subject of Approximations and Approximate Computation. It covers the analysis of the nature of approximate numbers, the analysis of numerical operations, the examination of the processes that arise in solving equations, the development of the proper technique of working with approximate data secured by means of measurement, and finally, the analysis of logarithmic and trigonometric functions. Part II is also designed as background material for the development of procedures 3 and 4, but it should be understood that the material presented in Part II is not on the secondary school level, although the Theory of Numerical Processes covers the entire field of secondary school mathematics.

The development of procedures 3 and 4 above was based wholly on the recommendations of the National Committee on Mathematical Requirements.

II. THE ANALYSIS OF AMERICAN TEXTBOOKS

From 1923 to 1932 eighty-eight algebra textbooks and thirty-four trigonometry textbooks were published in this country. Of these, only fourteen algebras and one trigonometry treat Approximate Computation in one way or another. The treatment of this subject in the one trigonometry is purely "experimental," and although much space is taken up, the exposition of the methods of Approximate Computation is entirely ineffective. Only two of the fourteen algebras have distinct units devoted to Approximate Computation.

In only eight algebras is the idea of measurement introduced. The idea of approximate numbers is thus directly tied up with measurement. There is in this respect very little deviation from the recommendations of the National Committee on Mathematical Requirements. Only one of the eight texts makes mention of the idea of an error in measurement. We shall refer to this textbook as Text A.[9]

The nature of approximate numbers is not clearly developed in any of the fourteen algebra textbooks. In none of them is distinction made between significant digits and decimal place approximativeness. In some these ideas are so intertwined and used so indiscriminately that the reader is confused.

The rules for rounding off digits are not standardized in any of the fourteen textbooks. In only seven are these rules directly mentioned. There exists no agreement as to the procedure in the case when the digit to be dropped is 5.

The case of addition and subtraction of approximate numbers is treated in all fourteen textbooks in a similar manner, that is, it is recommended that the approximate numbers be rounded off to the same number of decimal places.[10] The fact, however, that the last place of significance on the extreme right of the resultant sum or difference may very often cease to be significant is mentioned in only one textbook (Text B).[11] This fact could not have escaped the attention of the authors had they analyzed in detail the processes of addition and subtraction.

---

[9] The textbooks discussed in this study will be designated by letters.

[10] The wording "to the same place of significance" is more correct and more appropriate.

[11] The author, a teacher of considerable experience, makes the vague statement that in the case of addition and subtraction "if we add four approximate numbers, each of which is correct to 2 decimal places, the error in each may be as large as 0.005 and the maximum error in the sum is 4 times 0.005 or 0.02. The error in the difference of two approximate numbers of that same type may be as large as 2 times 0.005 or 0.01, because the two errors may lie in opposite directions." The author seems to be inconsistent in making this state-

The processes of multiplication and division were given scant attention by the authors of the algebra textbooks. In only one book, a ninth year algebra, did the author go to any length to discuss the process of multiplication and to give definite rules concerning the number of significant digits to be retained in the product. The rules given by this author are not, however, entirely correct.[12] The case of a product of three or more factors is not mentioned in any of the texts. The author of Text B becomes so confused that his rule for multiplication is more or less a mystery. He starts with significant digits, branches off to decimal places, and finally reverts to significant digits. He states his rule for multiplication (and division) as follows: "If two numbers to be multiplied do not contain the same number of significant figures, we may to save labor equalize the number of significant figures in the two given numbers. Thus instead of multiplying the approximate numbers 58.67 and 28.6, we may use 58.7 and 28.6, because in this approximate product we are sure only of the first two figures and for the third take one nearest to accuracy as determined by the figures following it." This statement is certainly not a rule in the ordinary sense of that term because it is not workable.

In not one of the fourteen algebra textbooks is there a word concerning either raising to a power or extraction of roots of approximate numbers.

Logarithmic and trigonometric functions, when treated in the algebra and trigonometry textbooks examined in this study, are referred to as numbers. In not one instance is there a statement that the approximativeness of the numerical values of these functions is in some relation to the approximativeness of the arguments of these functions. Neither in the algebra textbooks in which chapters on numerical trigonometry are found, nor in the trigonometry textbooks is there a word with reference to the evaluation of the approximativeness of the numerical results of the solved problems.

---

ment, because on one of the preceding pages he elucidates the idea of "compensation of errors," when the processes of addition and subtraction take place, and here he discards this idea and uses the process of obtaining the sum of the absolute values of the errors.

[12] This author suggests that two approximate numbers, whose product is to be obtained, should have the same number of significant digits, and that the product of these two approximate numbers should contain the same number of significant digits. The correct rule for multiplication appears on page 57.

In not one of the textbooks is there a word concerning the distinction between the idea of *precision* and the idea of *accuracy* of a measurement nor is anything said about the connection between these two ideas and the modes of representation of approximate numbers.

In all the textbooks the idea of accuracy of a measurement is confused with the idea of *correctness* in numerical computation. The word "accurate," if used when reference is made to a measurement, to a number, or to a result of a numerical process, is connected with the relative error of an approximate number. If reference is made to a *perfect* numerical result, the preferable word to be used is "*correct.*"

Of the fourteen algebra textbooks mentioned above only two have separate chapters on Approximate Computation. These are Text A and Text B. Text A has a unit on Approximate Computation in which the author discusses some aspects of the subject. The reader is first introduced to the idea of "digits we are certain of." This is connected with the exposition of the idea of measurement. The reader is then confronted with the notion of "significant digits." This notion does not simplify the situation because at first the reader is told that the "digits we are certain of" give some idea of the nature of the performed measurement, and then he is told that he may not be certain of the last digit (significant?) on the right. If the digit is significant there cannot be any question about it; but if the author raises the question of the certainty of this digit, he should modify his notion of significant digits. The author is silent on this point. He becomes entangled in an attempt to give two meanings for the same thing ("digits we are certain of" and "significant digits").

The rules for Approximate Computation in Text A are developed by means of the errors, both the apparent error for the case of addition and subtraction and the relative error (as well as the apparent error) for the case of multiplication and division. The respective errors for a sum, a difference, a product, and a quotient are derived in a general form, but no reference is made to the significant digits to be retained in a numerical result. There is no evidence of any kind that use has been made of the rules for the expressions for the errors given in Text A in the subject matter treatment.

Text B introduces approximate numbers without ever mentioning the fact that all measurements are represented by numerical values that are approximate. The rules for multiplication and division given in this text were discussed above. As far as the rules for addition and subtraction are concerned the author states that "no general rules can be given for the number of significant figures to be retained," and makes reference to errors which previously have not even been mentioned anywhere.

In conclusion, none of the textbooks examined treats the subject of Approximate Computation and Approximations satisfactorily. None of the texts carried out the recommendations of the National Committee on Mathematical Requirements so far as these two topics are concerned.

III. THE STATUS OF APPROXIMATE COMPUTATION IN FOREIGN COUNTRIES

A. SOVIET RUSSIA

The 1933 course of study for teacher-training institutions of Soviet Russia provides for an extensive course in Approximate Computation for teachers of mathematics and science. It should be understood that this course of study is not arbitrarily developed by one of the many Pedagogical Institutes of Soviet Russia, but bears the official stamp of approval of the Commissariat of Education and Public Instruction of the Soviet Union. In other words, the course is uniform for the entire country. From the material collected and examined it is evident that such a course has been given in the Pedagogical Institute in Twer for several years.[13] An advanced textbook based on a course in Approximate Computation given at the University of Leningrad has also recently been published,[14] as well as a series of very satisfactory textbooks for teaching Approximate Computation in the schools of all levels. Some of the latter have been republished several times.[15]

The students in the Pedagogical Institutes of the Soviet Union begin the study of Approximate Computation in their freshman year. The course is allotted ninety hours a semester. The total course covers the following topics:

---

[13] Bradis, V.: Teoria i praktika vychislenij (Theory and Practice of Computations); Part I, Arithmetic of approximate computations. State Educational and Pedagogical Publications, Moscow, 1933. This text is recommended by the Collegium of the Commissariat of Education and Public Instruction of the Soviet Union as a textbook for the Pedagogical Institutes of that country.

[14] Besikovitch, J.: Priblizhonnyje vychislenija (Approximate Computations). State Technico-Theoretical Publications, Moscow, 1933.

[15] Bradis, V.: Kak nado vychisliatj priblizhonnyje vychislenija (How to Perform Approximate Computations), in 3 parts. A textbook for secondary schools. State Educational and Pedagogical Publications, Moscow, 1932–1934.

Frank, M. L.: Elementarnyje priblizhonnyje vychislenija (Elementary Approximate Computations). A textbook for secondary schools graduates. State Technico-Theoretical Publications, Moscow, 1933.

A list of works on Approximate Computation published in the Soviet Union is found in the Bibliography (Appendix 4).

Numerical processes: addition, subtraction, multiplication, division, raising to a power, and extraction of roots.
The use and construction of tables (logarithmic, trigonometric, squares, and square roots).
The slide rule.
The use of computing machines.
Finite differences and interpolation.
The method of least squares.
The approximate solution of equations.
Nomography.

In the senior year the same course is repeated with a more thorough analysis of the above topics.

A university course in Approximate Computation is still wider in scope. Besides the above topics it covers:

Advanced theory of the approximate solution of equations.
Numerical differentiation and integration.
Graphical and mechanical integration.
Numerical integration of differential equations.
The harmonic analysis.
Advanced topics in interpolation.

B. GERMANY

There is no direct evidence of a definite requirement concerning Approximate Computation in the training of teachers, except that several textbooks based on courses given have been published. Such are:

Lüroth, J.: Vorlesungen über numerishes Rechnen. Leipzig, 1900.
Biermann, O.: Vorlesungen über mathematische Näherungsmethoden. Braunschweig, 1905.
Schrutka, L.: Zahlenrechnen. Leipzig, 1923.

None of these texts, however, devotes much space to numerical processes. The topics emphasized in these texts are concerned with such advanced phases of Approximate Computation as interpolation, and numerical differentiation and integration.

C. FRANCE

The prospective teachers of mathematics in France are expected to have a good knowledge of Approximate Computation. In the program for the Concours (examination) for 1914, the program of the "Lessons"[16] reads:

> Definition and extraction of the square root of a whole number or of a fraction to a given decimal approximation. Definition of absolute error and of relative error made in a sum, a difference, a product, a quotient, knowing the upper limits of the errors by which the given quantities are affected.[17]

A list of the texts on Approximate Computation published in France appears in the Bibliography (Appendix 4).

D. GREAT BRITAIN

Nothing has been done in Great Britain in regard to Approximate Computation, either in relation to the academic preparation and professional training of teachers or in connection with the teaching of this subject in the secondary schools. As late as 1932 the following statement appeared in one of the leading magazines in the field of mathematics in England:

> The reports of London University on the General School Examinations throw an interesting light on the knowledge of the candidates of the subject of approximations. The 1928-1930 report makes it quite clear that the subject is badly taught and that glaring errors occur very frequently.[18]

---

[16] Bulletin administratif du ministère de l'instruction publique, année 1913, 26 Julliet, pp. 175-177.

[17] Archibald, R. G.: The Training of Teachers of Mathematics of the Secondary Schools of the Countries Represented in the International Commission on the Teaching of Mathematics, p. 259. Department of Interior, Bureau of Education, Bulletin 1917, No. 27. Washington, Government Printing Office, 1918.

[18] Inman, S.: What Is Wrong with the Teaching of Approximate Numbers? The Mathematical Gazette, Vol. 16, pp. 306-315. London, 1932.

CHAPTER II

APPROXIMATE COMPUTATION IN SECONDARY SCHOOL MATHEMATICS

THE Report of the National Committee on Mathematical Requirements gave only general suggestions concerning the inclusion of Approximate Computation in the course of secondary school mathematics, and perhaps this fact accounts for the unsatisfactory state of the teaching of this subject. On the other hand, the dearth of theoretical literature thoroughly covering the field of Approximate Computation is another reason why Approximate Computation has made no headway in the secondary schools of this country.

No doubt, the stressing of the idea of the "exercise of common sense and judgment in computing from approximate data"[1] was misleading. The use of common sense and judgment, where definite rules might have been worked out for definite procedure, introduces an element of guessing. This suggestion may have influenced textbook writers to disregard the fact that with sufficient effort rules could have been worked out.

The National Committee recommended that junior high school pupils develop "an understanding of what is meant by the degree of precision as expressed by the number of significant figures."[2] This is the criterion not for the degree of precision but for that of accuracy.[3] The degree of precision is determined by the position of the first significant digit on the left of the number representing the apparent error of an approximate number.

The National Committee also recommended that pupils should become familiar with the effect of small errors on the results of the performed numerical processes. No attention was paid to this recommendation by any of the textbook writers (even by those who discussed "measurement") previously mentioned.

No mention is made in the National Committee Report of the fact that the solution of certain equations leads to approximate numbers.[4] The fact that many mathematical formulas, though mathematically exact, lead to approximate numerical values[5] is also disregarded in the Report. Nor does the Report mention the fact that the logarithmic and trigonometric functions are numerically represented by values which are generally approximate.[6] This last fact should change the fundamental aspect of the teaching of logarithms and of trigonometry (numerical trigonometry included) so that the meaning of the approximativeness of the numerical results of the operations with these functions will be emphasized to the pupils.

---

[1] The Reorganization of Mathematics in Secondary Education, pp. 7, 28, 50. A Report by the National Committee on Mathematical Requirements. The Mathematical Association of America, Inc., 1923.

[2] *Ibid.*, p. 22.

[3] See Part II, pp. 128 ff.

The writing of a secondary school textbook faithfully following all the suggestions about teaching Approximate Computation would not bring any nearer the realization of the recommendations of the National Committee on Mathematical Requirements. In order that a given discipline be properly taught it is imperative that those who write the textbooks as well as those who perform the act of teaching be thoroughly versed in the subject they are about to present. The only way to effect the belated introduction of Approximate Computation into secondary school mathematics is to provide all teachers of mathematics with an adequate foundation for this theory, especially within the limits of the Theory of Numerical Processes.

The practical numerical processes, that is, the actual performance of Approximate Computation, should follow the rules as developed in the Theory of Numerical Processes. There is no way to avoid the fact that the approximativeness of a numerical result of operations performed with approximate numerical data depends on the approximativeness of the given data as well as on the nature of the performed numerical processes. This dependence is extremely important not only from the purely mathematical point of view, but from the point of view of dependence in its relation to the pedagogical aspects of the teaching of mathematics. This dependence is of a certain limiting nature. The student of secondary school mathematics should learn to see for himself that he never can secure more information than is originally given to him, and that in many cases, by virtue of the performed numerical processes, he is bound to sustain a loss of factual information concerning the approximativeness of his numerical result. It is the purpose of Approximate Computation to give the pupil the necessary skill and to develop within him a mental attitude by means of which he can secure the optimum approximativeness without ever

[4] Part II, pp. 200 ff. [5] Part II, pp. 113 ff. [6] Part II, p. 233 and pp. 256 ff.

transgressing the limitations set down by the data, rules, and formulas provided.

The presentation of the subject of Approximate Computation in the secondary school in conjunction with the various other mathematical subjects taught is not an easy task. It is usual for pupils who begin to study secondary school mathematics to think in terms of exact numbers. The introduction of the notion of approximate numbers, that is, of numbers that actually should become more common in their experience, is not an easy matter. In everyday life a growing pupil, as his experience is enriched, comes in contact more and more with approximate numbers. He makes use of approximate numbers almost daily. He reads them in newspapers (baseball averages, averages of other kinds, prices); he uses them in purchases, and in measurements. All such activities require the development of an intelligent attitude toward approximate numbers. In many cases it is not only advisable, but necessary, that the attitude of pupils toward approximate numbers be intelligent. The pupils should learn to use these numbers carefully, that is, they should learn to exercise caution and never overestimate the importance of such numbers. They should always remember that approximate numbers do not possess some of the properties of exact numbers, that with approximate numbers in most of the cases one never can take out more information than he has put in.

In teaching the representation of approximate numbers one should emphasize the idea of a measurement. The numerical representation of a measure can be connected with the notion of measurement only by stating an approximate number in one of these three ways: (1) as correct to a certain unit (or units) of measurement, that is, hundreds, tens, hundredths, thousandths, and so on [illegible] as correct within a stated error of measurement; (3) as correct to a certain number of significant digits. The so-called decimal place approximativeness is of very little value in Approximate Computation either from the point of view of conveying direct information or from the point of view of the practical numerical processes, and it should never be used.[7]

In algebra, geometry, and trigonometry, whenever formulas are used, the pupils should be able to differentiate between "exact mathematical formulas," "approximate mathematical formulas," and "scientific formulas."[8] Formulas cannot perform miracles. Formulas are merely statements of relations between elements, and from the

[7] Part II, p. 120.

[8] Part II, pp. 113 ff.

point of view of practical numerical processes formulas are rules for these processes. At no time can a formula give a better approximativeness than its own nature and (or) the nature of the data involved. The pupils should make a practice of developing a critical attitude toward a formula so that they will always be conscious of the limitations which govern the approximativeness of their numerical results.[9]

It is deplorable that little, if any, attention is given in secondary school mathematics textbooks to the systems of units of measure and their relation to one another. Many secondary school pupils are not familiar with ways of expressing units of measure in terms of other systems of measure. Such information is instructive and useful. Experience in learning these systems would broaden the intellectual horizon of the pupils. At the same time the introduction of such topics into the mathematics classroom offers an excellent opportunity for the pupils to learn much about Approximate Computation and to do a considerable amount of practical work in this subject.[10]

There is veiled reference to two distinct problems in Approximate Computation in the Report of the National Committee on Mathematical Requirements.[11] These problems are stated in the Theory of Numerical Processes as follows:

Approximate Computation of the first type: The nature of the numerical data is definitely given, and it is required to apply the formulas and to perform the numerical computations in such a manner that the *optimum* approximation possible is obtained after all the stated computations are performed.

Approximate Computation of the second type: The nature of the numerical data is definitely given, and it is required to apply the formulas and to perform the numerical computations in such a manner that a result with a definitely *predetermined* approximativeness is obtained after all the stated computations are performed.[12]

None of the one hundred and twenty-two texts examined contains anything that would give even a hint as to the nature of the second type of Approximate Computation. Adequate attention should be given in the mathematics classroom to the case of Approximate Computation of the second type as well as of the first.

---

[9] Part II, pp. 193 ff.

[10] Part II, p. 118.

[11] See *(b)*, p. 3, above.

[12] Part II, p. 133.

CHAPTER III

DEVELOPMENT OF APPROXIMATE COMPUTATION IN SECONDARY SCHOOL MATHEMATICS

The subject matter presentation of Approximations and Approximate Computation in the secondary schools should be made on a limited scale and at the same time on the level of difficulty of the mathematical subjects presented to the pupils. It should be understood that the writer does not recommend the presentation of the subject of Approximations and Approximate Computation as a separate topic, nor does he recommend the spreading of this topic throughout the entire course of secondary school mathematics. The method of procedure in this case depends on the textbook writer and the individual teacher. It should be understood, however, that in the light of the recommendations of the National Committee on Mathematical Requirements the ideas of Approximations and Approximate Computation ideally should permeate the entire course of mathematical instruction in the secondary schools.

The aim in presenting the subject of Approximations and Approximate Computation in elementary algebra and numerical trigonometry is to introduce the notion of measurement and to develop the concept of an approximate number as the numerical value of a performed measurement as well as simple rules for working with approximate data.

Whenever the derivation of a formula for Approximate Computation is given in the material in the following pages, this derivation is based on the more rigorous treatment developed in Part II.

The subject of Approximations and Approximate Computation in elementary algebra and numerical trigonometry is divided into seven major topics:

I. Measurement.
II. Approximate numbers.
III. Numerical operations.
IV. Logarithms.
V. Natural trigonometric functions.
VI. Approximate Computation of the first type.
VII. Approximate Computation of the second type.

The suggested development of the subject matter of each topic is presented in outline form. Each section and unit of this outline is supplemented by typical illustrative material and extensive teaching remarks devoted to the analysis of the outstanding features of the respective sections and units. In order to avoid unnecessary repetition, the following abbreviations will be used: *I.P.M.*, Illustrative problem material. *T.R.*, Teaching remarks.

The development of the processes of Approximate Computation on the level of intermediate algebra presupposes the familiarity with the fundamental notions of Approximations and Approximate Computation developed on the elementary algebra level. In other words, it is expected that whatever concepts of measurement and its approximativeness have been developed in a course preceding an intermediate algebra course, these should be retained and carried along by the pupils. Therefore, the teacher should not interpret the development of the methods of Approximate Computation on the intermediate algebra level presented here as being purely mechanical. Generally speaking, the technique of Approximate Computation is not mechanical. The performance of numerical processes may be thought of as mechanical, but the arithmetic of Approximate Computation can be fully appreciated if and only if the interpretative processes are predominant. Only when a pupil is conscious of the nature of his data and can interpret the approximativeness and the meaning of the numerical results obtained by him, will he understand the importance of Approximate Computation as a fundamental part of applied mathematics.

The above remarks refer also to the development of the processes of Approximate Computation on the level of advanced algebra and trigonometry. Again, here in this work, only numerical processes are considered. The applications of Approximate Computation to such topics as statistics, business mathematics, and science can scarcely be avoided on the advanced algebra level. The teacher as well as the textbook writer should not fail to recognize the proper topics to which Approximate Computation should be applied, and should not miss the opportunity of making thereby their presentation more meaningful. But whatever is the place allotted to Approximate Computation in secondary school mathematics, Approximate Computation should not be treated as a localized topic. Approximate Computation is not a topic—it is a method of procedure in mathematics, and it should be taught as such.

CHAPTER IV

APPROXIMATE COMPUTATION IN ELEMENTARY ALGEBRA AND NUMERICAL TRIGONOMETRY: MEASUREMENT

THE purpose of this chapter is to develop the idea of measurement as a necessary background for the introduction of approximate numbers. Besides a discussion of this topic, it contains much informational material which has considerable practical and cultural value. Though at first the topic may appear to contain a preponderance of reading material, this impression is not altogether correct. The teacher should devise a suitable activity, and a resourceful teacher can use the topic for reviewing arithmetical work in the beginning of an algebra course.

I. THE DEVELOPMENT OF THE NOTION OF MEASUREMENT

A. THE NEED OF MEASUREMENT AS A MEANS FOR COMPARISON OF QUANTITATIVE AS WELL AS QUALITATIVE PROPERTIES OF THINGS

I.P.M. The uses of measurement in human activities such as trade and commerce, governmental work, play, school work, and science.

T.R. The selection of the illustrative material as well as the development of this unit entirely depends on the resourcefulness of the teacher. The introduction of the notion of measurement may make the best appeal if and when the evolution of measurements is presented historically. The pupil can learn more by going through the process the race has gone through. Considerable amount of illustrative material for classroom activity purposes may be secured from sources such as Dr. Smith's History of Mathematics (Ginn and Company, Boston, 1926), and articles from the Encyclopaedia Britannica, eleventh edition, on calendar, clock, mensuration, money, time, weighing machines, weights and measurements. At the end of each of these articles there is an extensive bibliography which may be utilized to enlarge and extend the material a teacher may need in developing and illuminating the subject.

B. THE KINDS OF MEASURING INSTRUMENTS

I.P.M. The practical use of ruled tape, weighing scales, clock, protractor, and simple physical instruments such as graduated glassware, precision scales, and the like.

T.R. Most of these measuring instruments are easily obtainable and should be a part of the permanent equipment of a mathematics classroom, not only for exhibit purposes but for practical work. The use of these instruments will enrich the mathematics course.

C. THE UNITS OF MEASURE

I.P.M. The development of the idea of a *standard unit* of measure.

T.R. The idea of standard units of measure should be developed from the point of view of a general agreement made by a large body of men. The evolution of units of measure is an interesting study, which can be utilized to motivate material for many problems, and to develop the idea of functionality, especially in connection with standardized units which were originally determined arbitrarily; for example, the yard and the barleycorn. See references given in sec. I, A, above. This unit offers opportunity for considerable numerical work, especially review work in arithmetic.

i. Length

I.P.M. The English and metric systems of measures of length. References to the historical development of the standard units of measure of length now in use. The advantages of the metric system. The translation of numerical values of length from one system of measurement into another.

T.R. When introducing the English measures it should be pointed out to the pupils that the yard was originally an arbitrary measure determined by King Henry I. Such questions as the use of another standard than the length of the king's arm will elucidate the idea of functionality.

In contrast to the English system the metric system, which is based on geodetic and physical measurements, offers opportunity to stress the stability of this system (page 118).

ii. Area

I.P.M. The idea of the unit for the measure of area as the square of a unit of measure of length. The development of the units of measure of area in the English and metric systems. The translation of measures from one system into another.

T.R. The development of the notion of the unit of measure for an area should be built up from the unit area of a square. The derivation of the unit of measure for an area in a given system of measures and its relation to some higher units of measure of length in the same system should be worked out by the pupils as an exercise. Similarly the relationship between units of measure of area in the English and metric systems should be based on the relationship between the linear measures in these systems, and it should be worked out as an exercise.

iii. Volume

I.P.M. The idea of the unit of measure of volume as the cube of a unit of measure of length. The development of the units of measure for volume in the English and metric systems. The translation of measures from one system into another.

T.R. Whatever remarks have been made in regard to the units of measure for areas are applicable to the units of measure for volumes.

iv. Weight

I.P.M. Historical illustrations of the evolution of the measures of weight. The English and metric systems of measures of weight. The standardization of the units of measure for weight. Physical constants of volumes of water for units of weight. Practical applications such as the use of weight to denote volume (tonnage).

T.R. The development of the systems of measures of weight should follow the same genetic process employed in the development of the units of measure for length, area, and volume. It is worth while to connect the volume of water with units of measure of weight. This will stress the fact that units of measure of weight are generally not arbitrary. Moreover, in shipping much use is made of this procedure when the size of ships is expressed in tonnage.

v. Content

I.P.M. The English and metric units of measure for content. The translation of a unit of measure of one system into a unit of measure (or units of measure) of another system.

T.R. The units of measure for content should be connected with both the units of measure for volume and the units of measure for weight.

vi. Angles

I.P.M. The sexagesimal and centesimal (optional) systems of units of angle measure.

T.R. When introducing the units of angle measure the teacher should not fail to mention the motivating factors in the use of angle measurements in engineering, in science, in time reckoning, and in navigation. Some pertinent historical remarks on the origin of the sexagesimal system will prove interesting when time measures are studied.

vii. Time

I.P.M. The calendar. Some historical illustrations of the evolution of the calendar. The idea of B.C. and A.D. The units of measure of time. Historical illustration of the development of the sexagesimal system in measurement of time.

The clock. The earth as a clock. The instability of all clocks. The idea of synchronization.

The year. The leap year and its causes.

T.R. The material on the calendar is very interesting and rich with examples that lend themselves for problem developments and computational work, such as the relation between the Mohammedan year and the year as reckoned by Christian people. The time reckoning systems of various peoples are exceedingly instructive, and at the same time a classroom discussion of these topics offers an opportunity to make clear the relation between these systems and the directed number system.

The teacher should point out that no clock is perfect, and even though our time reckoning is based on the revolution of the earth on its axis and the motion of the earth around the sun, the earth is an imperfect clock because these two motions are not regular.

viii. Money

I.P.M. The evolution of the United States dollar. Systems of money in foreign countries. The practical significance of the decimal system. The translation of units of money of one country into units of money of another country.

T.R. When teaching the systems of units of money use can be made of daily quotations of foreign exchange. These quotations offer excellent material for problems with arithmetical work. When teaching the United States units of money the teacher should consult F.

Cajori's History of Mathematical Notations (Open Court Publishing Co., Chicago, Ill., 1929) especially in regard to the evolution of the signs for the dollar.

ix. Scientific units of measure such as in meteorology, astronomy (the time-distance), electricity, and sound

I.P.M. The temperature. The average rainfall. The light year. The distance to stars. The translation of light years into miles and kilometers. The translation of the distances to the sun and to the moon in terms of time. The watt and the kilowatt. The bel. The angström. The cycle and kilocycle.

T.R. The introduction of the "average rainfall" as a unit of measure should eventually lead to a thorough development of the notion of an average. This notion will be developed later as an example of an indirect measurement.

The units of measure such as the light year, the degree of temperature, and the electrical units should be developed in the classroom with sufficient illustrative problems showing their uses. These units of measure as well as many others (the radio wave lengths and their frequencies) may lead to very interesting problems. For example, a teacher may develop a unit on "Mathematics and Music" which may be presented on the level of elementary algebra. Radio wave lengths and kilocycles are household words, and every one uses them, not knowing, however, what they measure.

II. The Process of Measurement

A. DIRECT MEASUREMENT AS A PROCESS OF COMPARISON OF THE MEASURED PROPERTY OF AN OBJECT WITH A STANDARD UNIT OF MEASURE

I.P.M. Classes of objects possessing the same property, such as the class of tables, the class of chairs, the class of books, the class of clothing. The idea of a subclass, such as round tables, long tables, square tables. The development of the idea of a member of a class as a unit, thus leading up to numbers and counting.

T.R. The development of the idea of a class as a totality of objects possessing a common property, each member of this class being a discrete unit, is the most logical procedure by means of which the idea of a number can be built up. After the idea of a class is thoroughly developed the idea of a subclass logically follows therefrom.

This method easily leads to the introduction of algebraic symbolism by means of which a letter represents the name of the class, and the number of the units (or members of the class) is represented by the coefficient of that letter. Thus 50 books, of which 10 have green covers, 10 have red covers, 18 have blue covers, and 12 have yellow covers, can be represented symbolically as follows:

$$50B = 10g + 10r + 18b + 12y.$$

i. Counting—exact numbers (pages 113, 117)

I.P.M. The number of pupils in a classroom. The number of pages in a book. The number of seconds in a minute, in an hour, in a day. The number of feet or yards in a mile.

T.R. The process of counting is the simplest kind of measuring. The count of a certain number of objects is a measure, and this represents the first step in measurement taken by the race.

ii. The application of a measuring instrument to the object that is measured

I.P.M. Practical work in measuring lengths, widths, and heights of objects. The measurement of weights of objects. The measurement of angles with a protractor.

T.R. The use of practical work in measurement in a classroom will make the problem of measurement real to the pupil, and it will help to create the proper background for the introduction of the concept of approximate numbers which represent the numerical values of performed measurements (except counting, which is represented by exact numbers). Moreover, practical work in measurement will familiarize the pupils with the processes they will have to make use of very often in their lives.

iii. The illustration of the impossibility of securing an exact measurement (pages 117, 118)

I.P.M. Repetition of the same measurement several times by one or several pupils in order to demonstrate that no measurement, even though repeated several times, can be satisfactorily stated in terms of the same numerical value. A discussion that would lead to the discovery of the reasons for these discrepancies. Comparison of several instruments of measure of the same kind. The illustration of the so-called *personal* defects of an individual performing a measurement, such as inability to see well or to estimate measures.

T.R. The pupils should be led to see for themselves that measurement is not a simple process, that measurement involves certain factors which cause slight deviations in the numerical statements of the performed measurements. The discovery of these factors is very important in building up the proper concept of approximate numbers.

iv. The numerical value of a measurement is expressed only approximately (pages 113–121)

I.P.M. The agreement to use the numerical value of a measure to the nearest known mark (or division) on a measuring scale. Practical examples illustrating this agreement. The fact that all measures secured in such a manner are approximate.

T.R. Actually this agreement signifies that the numerical value of a performed measurement, which we denote by a, is such that

$$A_2 < a < A_1,$$

or, for example,

$$156 \text{ in.} < a < 157 \text{ in.}$$

Then, if a is nearer to the 156th division than to the 157th division we agree to call the measure 156 in. If a is nearer to the 157th division than to the 156th division we agree to call the measure 157 in.

v. The obtained measurement is located between two units known exactly and given by exact numbers (pages 123–125)

I.P.M. Illustrations showing that the agreed method of stating a measure involves an exact number of units which can be counted. The two measures between which the desired measure is located are known as the upper and the lower bounds of this measure.

T.R. Elaborating on the statement $A_2 < a < A_1$ it can readily be seen that A_2 and A_1 contain a total number of units to which the measurement was taken, and these units can be easily counted. Since they can be counted, and if no mistake is made in counting, A_2 and A_1 are represented by exact numbers.

vi. The error in measurement. The idea of precision in measurement. The error in measurement is known as the apparent error (pages 122–125)

I.P.M. Development of the idea that the *amount* left over and not counted in a measure is still a part of the measure, and should be taken into consideration. Illustration of the fact that this amount varies in size for different measures expressed in terms of the same unit of measure. This amount is known as an *error* in measurement, and is less than a single unit of measure in terms of which a measurement is stated. Introduction of the expression in measurement "correct to inches," "correct to tenths of a foot," "correct to one inch," and the like. Illustrations showing that the smaller the unit of a measure in terms of which a measurement is stated, the more precise is the measurement. Comparison of various instruments of similar kind in order to determine which is the more precise.

T.R. The idea of error in measurement is exceedingly important because the analysis of the error enables us to judge the quality of the measurements. The pupils should be shown that an error in measurement is always less than a single unit in terms of which a measurement is stated. If the error were a unit or greater than a unit in terms of which the measurement is stated we could not be certain of that unit, and the measurement should be considered as having been performed to a unit greater in size. This should lead up to the idea of the precision of a measurement. This notion of the relation between the magnitude of an error in measurement and the magnitude of the unit in terms of which a measurement is stated is important in connection with the notion of significant digits which will be developed presently. The evaluation of the error in measurement should be postponed, however, until the mathematical properties of the arithmetic mean are discussed (see sec. II, B, *i* and *ii*, page 27). The expression "apparent error" was found by the writer to be most commonly used by some European writers in the field of Approximate Computation. This expression may be supplanted, however, by the expression "possible error" which from some points of view may prove to be more popular among teachers and textbook writers. The use of the word "residue" or the word "residual" may be also suggested. These terms are consistent with the meaning of an apparent error of an approximate number; that is, a residue (or residual) is what remains after a certain portion of an exact value (which is not definitely known) is determined, and the remaining portion (or its upper limit) is taken as the error.

B. INDIRECT MEASUREMENTS REPRESENTING RESULTS OF OPERATIONS WITH NUMERICAL VALUES OF PERFORMED MEASUREMENTS

I.P.M. The use of indirect measurement representing computation which is performed with numerical values of measures already secured. The computation of the total weekly savings of a class. Multiplication problems which lead to the determination of the cost of several objects of the same kind when the price of one object is given.

T.R. The development of the material in this section should aim to show that many measures are obtained by means of computation which is performed with the numerical values of measures already secured. This development should aim to motivate the use of rules for computation with approximate numbers.

i. The average as a measure (page 112)

I.P.M. The average grade of a pupil. The average rainfall. The average height or weight of the pupils in a class. The average attendance of a class in a given period. The fact that an average may be represented by fractional parts of individual units. For example, the size of the average family in New York State according to the 1930 census is given in the World Almanac as 3.63 persons. The formula for the average; the arithmetic mean.

T.R. The idea of an average should be developed in terms of the arithmetic mean. A teacher can find many illustrations and illustrative examples for the derivation of the arithmetic mean in many fields of human activity as well as in cases arising from actual experience of the pupils. The statistical significance of the arithmetic mean as a measure of central tendency is very interesting and instructive; this fact should be pointed out to the pupils. The teacher should explain that the notion of a central tendency means that in case a measurement is repeated a great many times, or a series of measures of the same object are given, the most probable numerical value of the measurement in question can be safely taken as the arithmetic mean of these measures.

ii. The error in measurement as the average of the numerical values of the deviations of the individual measures from their arithmetic mean. The direction of the error (pages 122–125)

I.P.M. Practical measurements of lengths, for example, repeated several times. The computation of the average (arithmetic mean) of these measures, and then of the deviations of the individual measures from their arithmetic mean. The calculation of the arithmetic mean of the numerical values (that is, disregarding their respective signs) of these deviations. (This arithmetic mean is defined as the error.) Illustration of the fact that the error should have the sign ±, because there are measures exceeding as well as measures that are less than their arithmetic mean. For example, the following measures were obtained: 25.4 in., 25.6 in., 25.3 in., 25.5 in., 25.7 in., 25.4 in., 25.5 in. Then the average measure is

$$\frac{25.4 + 25.6 + 25.3 + 25.5 + 25.7 + 25.4 + 25.5}{7} = 25.5 \text{ in.}$$

The arithmetic mean of the deviations is

$$\frac{0.1 + 0.1 + 0.2 + 0 + 0.2 + 0.1 + 0}{7} = 0.1$$

The apparent error is then ±0.1 in. The conclusion is that the measurement is correct to inches only.

T.R. The purpose of this unit is to make sure that the pupils learn the method of calculating errors. The calculation of errors is usually effected by means of the method of least squares, but the arithmetic mean of the absolute values of the deviations is suitable for the purpose of calculating the errors on the elementary algebra level because it is proved in the theory of least squares that the arithmetic mean of the absolute values of the deviations of the individual measures from their arithmetic mean is greater than the square root of the mean square deviation.

At this stage the teacher should not fail to develop the notion of the direction of the errors. The results secured in this unit are very important for the process of developing the notion of significant digits, especially from the point of view of the positional location of the first significant digit on the left of the apparent error.

iii. The measure of the area of a rectangle expressed as a product of two linear measures

I.P.M. The derivation of the formula for the area of a rectangle: $A = lw$. Actual measurements of the lengths and widths of rectangles with one system of units of measure (for example, English) and the translation of the numerical values of areas into metric units. Then actual measurement with metric units. A comparison of the numerical results. Do the results agree? What is the cause of a disagreement, if any?

T.R. The purpose of this unit is to show the need of a rule for multiplication of approximate numbers. The motivation can be effected by letting the pupils secure numerical results, as stated above, which would vary to a certain extent, although such a variation should not be expected if the numbers were exact. The teacher should point out that the value of an inch, for example, in terms of centimeters can be determined approximately only, hence the cause for a seeming, but unavoidable, discrepancy which necessitates the use of a rule for multiplication of approximate numbers.

iv. The measure of the volume of a rectangular solid expressed as a product of three linear measures

I.P.M. The derivation of the formula for the volume of a rectangular solid: $V = hlw$. Make use of the same procedure as in the case of the calculation of areas in sec. *iii* above.

T.R. See sec. *iii* above.

v. Ratios as measures

I.P.M. The illustration of the fact that all measures are ratios of the length, area, volume, weight, and so on, of an object to a given unit of measure.

A ratio of two quantities is a measure for the purpose of comparison of these two quantities. Hence a ratio is also a kind of measure.

Comparison of the heights or weights of two persons. Comparison of the savings of two pupils. The idea of percentage.

T.R. When developing this unit the teacher should stress the fact that the ratios represent measurements of relationship between measures. The teacher can make use of the notions of density and specific gravity from physics. The introduction of the notion of index numbers and their statistical significance will enable the teacher to use many problems from various fields of human activity. The proper development of this unit will save much work in the future when ratios are applied (similarity in geometry, numerical trigonometry, percentage problems).

In order to motivate the use of a rule for division of approximate numbers the teacher should use a procedure similar to those outlined in sec. *iii* and sec. *iv* above.

vi. Prices as measures

I.P.M. The determination of prices of one unit or of several units when the cost of a given number of units is given. In case of fractions of cents, the demonstration of the common procedure of expressing the price in cents. The fact that the price if expressed in cents is very often only approximate.

T.R. The teacher should point out to the pupils that in the case of money, although the price may be only approximate, the numerical statement of the cost or of the price given to the last unit of money-measure is in practice taken as being exact. Thus in money problems the processes with numbers obey the rules for exact numbers. The only exception would be in the case of the conversion of foreign exchange when the various equivalents are approximate.

vii. The need of rules for working with approximate numbers

I.P.M. Additional numerical problems illustrating the development given in sec. *i* to sec. *vi* above. The stressing of the necessity of securing definite rules of procedure for working with approximate numbers which represent the values of obtained measurements.

T.R. This unit is to serve the purpose of a review of the entire topic so that the pupils will be prepared for the introduction of approximate numbers and the analysis of their properties.

CHAPTER V

APPROXIMATE COMPUTATION: APPROXIMATE NUMBERS

The purpose of this chapter is to analyze the nature and the constituent properties of approximate numbers as well as the relation between the errors of approximate numbers and the qualitative properties of the secured measures, that is, their precision and accuracy. The development of the notions of precision and accuracy of approximate numbers should be made gradually so that each section of this topic will lead into the next one. Ample numerical illustrations will enable the teacher to avoid certain technical difficulties which will be pointed out whenever such arise.

I. The Origin and Kinds of Approximate Numbers

Since throughout this section more stress will be laid on the arithmetical nature of approximate numbers than on the practical side of the securing of this kind of numerical data, the teacher should try to recreate some of the illustrations previously presented to the pupils so that they will retain the connection between measurements and approximate numbers representing the values of the obtained measures. During the development of this as well as the succeeding topics, whatever is to be said in regard to approximate numbers should arise as a result of an intuitive (so to speak) study of measurement.

A. APPROXIMATE NUMBERS ARISING FROM PROCESSES OF DIRECT MEASUREMENTS

I.P.M. Review of illustrations presented in the process of the development of the topic "Measurement" in order to recreate the process of obtaining approximate numbers by means of direct measurements.

T.R. When discussing the nature of approximate numbers arising from processes of direct measurements, the teacher should not fail to emphasize the fact that approximate numbers so obtained relate directly to the precision of the measuring instrument with which the measurement was performed.

B. APPROXIMATE NUMBERS ARISING FROM PROCESSES OF NUMERICAL OPERATIONS (pages 112–119)

I.P.M. Problems in which repeating and (or) non-terminating decimals such as 0.3333 . . . , 0.6666 . . . , and 0.142857 . . . occur.

Problems involving the extraction of square roots when the number under the radical is not a perfect square, such as $\sqrt{2}$, $\sqrt{3}$, $\sqrt{5}$, and $\sqrt{7}$.

T.R. This unit introduces a series of numbers which are approximate not because they represent results of direct measurements but because their nature is determined by the type of the numerical processes that are involved during the calculation of these numbers. Usually these processes are division and the extraction of roots. This unit may offer some difficulties in conveying the notion of the approximativeness of such numbers because of the fact that for one reason or another a given number is taken to a desired number of significant digits, thus either exceeding the exact value of the number or being less than its exact value. The pupils should be taught to recognize the cases where such approximate numbers may arise so that they will become accustomed to treating such numbers as approximate. The teacher should explain to the pupils that such numbers as $\frac{5}{6}$, $\sqrt{2}$, $\sqrt{3}$, if taken as they are, are exact, but when their numerical values are expressed in the decimal number system, these numerical values can be obtained in the form of approximations only.

C. CERTAIN MATHEMATICAL CONSTANTS WHICH CANNOT BE REPRESENTED EXACTLY IN THE RATIONAL NUMBER SYSTEM

I.P.M. Problems involving the calculation of the circumference and (or) the area of a circle. The use of $\pi = 3.14159$ . . . taken to a definite number of significant digits. The fact that the numerical value of π cannot be represented exactly in the rational number system.

T.R. The need of the use of irrational numbers as $\sqrt{2}$, $\sqrt{3}$, or values of π, arises very early in secondary school mathematics. The introduction of a transcendental number such as π should be done with great care and caution. To many pupils the approximate value of $\pi = \frac{22}{7}$ is unfortunately presented as an exact value which is fundamentally wrong and, to put it mildly, misleading. It is better to introduce the numerical value of π as essentially approximate, and some historical illustrations of the attempts to calculate this value to many digits may prove to be helpful and convincing. At no time, however, should a pupil be given a chance to form the notion that π can be represented exactly in the rational number system.

II. THE ERROR IN MEASUREMENT AS A PART OF AN APPROXIMATE NUMBER (pages 112–125)

The material presented in this section should aim to stress the fact that all information necessary for the study of the nature of the numerical values of measurements, that is, approximate numbers, is obtained by means of the numerical values of the apparent errors. The development of this topic is designed to lead up to the notion of significant digits, and, as it will be pointed out presently, the notion of significant digits hinges on two ideas: the idea of a measurement and the idea of the positional location of digits in the decimal number system. The positional location of the first digit on the left of an apparent error, this digit not being zero, controls both the precision of a measurement and the positional location of the last significant digit on the extreme right of an approximate number.

A. THE APPARENT ERROR

I.P.M. Numerical illustrations showing the necessity of stating the error of a measurement which indicates the units to which the particular measure can be considered reliable, that is, correct. Thus (16.2 + 0.03) in. shows that the error is in the hundredths. The error is a part of an approximate number.

T.R. The teacher should point out this fact as a statement of general agreement. The necessity of knowing the apparent error is important in all the stages of Approximate Computation. The pupils should be led to see that if the apparent error is known the determination of the approximativeness of a numerical result becomes possible.

B. THE RELATION BETWEEN PRECISION AND THE POSITIONAL LOCATION OF THE FIRST DIGIT ON THE LEFT (NOT ZERO) OF AN APPARENT ERROR OF AN APPROXIMATE NUMBER

I.P.M. Further numerical illustrations showing that the precision of a measurement is determined by the positional location of the first digit on the left (not zero) of the apparent error. Comparison of different approximate numbers in order to determine the relative precisions of the measures they represent.

T.R. At this stage the teacher should introduce the idea of a positional location of a digit in any number. Use of this notion will be made very often hereafter, and the gradual development of this notation is essential for the introduction of the notion of significant digits. The teacher should demonstrate the reason why in the examination of apparent errors one should be interested in the first digit counting from the left which is not zero, that is, the digit representing a magnitude but not the absence of magnitude.

C. THE DIRECTION OF AN APPARENT ERROR

I.P.M. Graphical and numerical illustrations showing the location of a measurement between two divisions on a measuring scale. The fact that any measure can thus be stated as being taken either with an error in excess of a given division, that is, with a positive error, $(16.2 + 0.03)$, or with an error in defect of its following division, that is, with a negative error, $(16.3 - 0.07)$. Use of numerical values of $\sqrt{2}$, $\sqrt{3}$, and the like to illustrate the same property.

T.R. This unit is presented on the assumption that the pupils are familiar with the directed number scale as well as with the fact that the description of a measurement includes the direction of the measurement as well. In the experience of the pupils by this time, observation of the fact that a measurement is located between two units (divisions) on a ruled measuring scale should be a well-known step in the process of measurement. Thus the discussion of the material of this unit completes the analysis of the process of measurement. The use of squared paper and physical apparatus such as thermometers and weighing scales for practical work covering the material of this unit should prove very effective.

D. THE GREATEST POSSIBLE APPARENT ERROR IF IT IS NOT STATED DIRECTLY

I.P.M. Illustration of the agreement that an apparent error is assumed to be equal to and never greater than 0.5 of a unit in terms of which a measure is expressed, that is, to which the measure is correct.

T.R. When an approximate number is stated without its apparent error, the magnitude of the apparent error may be considered as having any of the following values: 0.1, 0.2, 0.3, 0.4, or 0.5 of a unit in terms of which a measure is stated. Should the magnitude of the apparent error of an approximate number be greater than 0.5 (that is, when it is so given) of a unit in terms of which the measure is stated, the representation of an approximate number may be reversed; that is, if the apparent error is positive, the approximate number may be rewritten so that it will have a negative apparent error. Thus $(16.2 + 0.07) \equiv (16.3 - 0.03)$. If the apparent error of an approximate number is negative, the approximate number may be rewritten so that it will have a positive apparent error. Thus $(16.2 - 0.07) \equiv (16.1 + 0.03)$. But if the apparent error is not definitely stated, then the assumption to be made should consider the most unfavorable case, that is, the greatest possible apparent error—0.5 of a unit in terms of which the measure is stated.

E. THE DIRECTION OF THE GREATEST POSSIBLE APPARENT ERROR IF THE APPARENT ERROR IS NOT STATED DIRECTLY

I.P.M. Use of approximate numbers when the apparent error is not given. The assumption that the apparent error is ±0.5 of a unit of measure in terms of which the measure is stated; that is, that the error lies in either direction. The agreement to make use of this type of apparent error throughout the entire work of elementary algebra and numerical trigonometry.

T.R. The determination of the direction of an apparent error of an approximate number when the magnitude of the apparent error is not given may involve some difficulties. The teacher should use ample numerical illustrations showing that when a numerical value of a measure is stated without the apparent error, the measure can represent a measurement obtained either in excess or in defect of the true value, and these two cases must be taken into consideration simultaneously. For example, if a measure is stated as 32.6 in., then it may be either at least as much as 32.55 in., or at most as much as 32.65 in. Hence the assumption for the direction of the error should be ±, and the approximate number may be written as (32.6 ± 0.05) in. The teacher should note that this mode of writing corresponds to the notion $32.55 < 32.6 < 32.65$. In the numerical processes of elementary algebra and numerical trigonometry we shall not be concerned with the actual magnitude of the apparent error. Hence for all approximate numbers, whether they are obtained from actual measurements or from tabular values of logarithmic or trigonometric functions, and whether they are natural or logarithmic, the apparent error will be assumed as ±0.5 of a unit of measure in terms of which the measurement is stated.

III. Significant Digits (pages 119–123)

The purpose of this section is to develop the notion of significant digits and to lead up to the rules for rounding off digits. The concept of significant digits is of great importance in Approximate Computation because it is directly related to the notion of *accuracy* (see sec. v below); and in connection with apparent errors, which determine the significance of the digits on the right of an approximate number, significant digits are indirectly related to the notion of precision. The concept of significant digits must, therefore, be carried through all the processes of Approximate Computation.

The expression, "the number of significant digits of an approximate number," may sound somewhat awkward. This difficulty arises from the fact that there is no other word in the English language that can be substituted for the word "number" as used the first time in the above expression. In German, for example, the word "number" as used the first time would be *Anzahl* (which represents the counting of the digits), and as used the second time, *Zahl*.

A. THE DECIMAL SYSTEM AND ITS POSITIONAL PROPERTIES

I.P.M. The use of numbers to illustrate the connection of the positional location of a digit in a given number with the relative magnitude of the unit in terms of which the digit is expressed. For example,

$$247 = 2 \cdot 100 + 4 \cdot 10 + 7.$$

The fact that each unit to the left of a certain unit is ten times as large as that unit.

T.R. The positional properties of the decimal system of numbers are seldom, if ever, presented to pupils of elementary algebra and numerical trigonometry. These positional properties are important not only from the theoretical point of view, as discussed in advanced stages of secondary school mathematics when various number systems are studied, but from the point of view of measurement they should be brought to the attention of the pupils as early as possible in their studies of elementary algebra. In reality the digits are not significant, but the units to which these digits correspond are important and therefore significant. The correct expression would be "significant units" or "significant places," and the units in terms of which the apparent error of a measurement or of an approximate number representing this measurement is expressed are not significant and are not important in the numerical value of this measurement.

When importance is attached to a given unit it is done because this unit is of value in the numerical statement of a measurement. The magnitude of the digits (or of the number of units) is of secondary importance. Primary consideration in this case is given to the question whether we should concern ourselves with a given unit or not. The expression "significant digits" is thus a result of a transfer of the concept of significant units to the digits representing the number of these discrete units. Actually when the expression "significant digits" is used, one should think of the units or places and not of the digits, and thus avoid confusion regarding the last digit on the extreme right of any approximate number. For example,

$$2.67 \pm 0.005,$$

actually is identical with the statement

$$2.665 < 2.67 < 2.675.$$

The hundredths units are important and therefore significant, and we say that the digit 7 is significant. This notion should be thoroughly developed because it will be used throughout the entire work in Approximate Computation.

There should be thorough classroom discussion of such questions as this: Suppose that in the number 247 the digits 4 and 2 interchange places. Which digit becomes the more important one? The fact is that 4 becomes the more important one because it then represents the hundreds units whereas 2 then represents the tens units only, and the former units are ten times as large as the latter ones. The pupils should see that the magnitude of the number represented by the digits is immaterial in such a case. This is a new and important concept, the concept of *order.*

B. THE SIGNIFICANCE OF DIGITS REPRESENTING A MEASURE

I.P.M. Illustrations of numerical values of measures demonstrating the same properties discussed in section A above. The development of the idea of significance of the digits on the basis of these illustrations. The connection between the value of a unit of measure and the place of significance.

T.R. After the concept of the positional properties of the decimal system is developed, illustrations of measures represented by numbers are appropriate. The notion of every digit representing a certain number of units of measure as occupying a place of significance and therefore of importance in the process of imparting information concerning the measurement should lead to the definition of significant digits.

C. THE DIGIT ZERO

I.P.M. The significance of zero (or of several zeros) if it is located between two significant digits, or if it is located on the right of a number and its significance follows from the statement of the measurement. The fact that the zeros on the left of a number, preceding the first significant digit (which is not zero), are not significant. For example, in 2.072 the zero is significant. In log 2 = 0.3010 the second and third zeros (counting from left to right) are significant, whereas the first zero is not significant.

T.R. The number zero (0) is a stumbling block for many pupils. Considerable attention should be given by the teacher to the development of the proper notion of the significance of the number zero, especially when this number signifies the fact that a measurement was performed but there were found no units of a given denomination in that measurement. Zero in this case does not stand for the concept "nothing"; zero signifies in this case the *absence* of certain units. If zero stands for "nothing" it is not significant, as, for example, in the decimal fraction 0.0034. The two zeros are significant in cases such as log 20 = 1.3010. There is a certain amount of difficulty involved in the case when a zero (or several zeros) is on the right of a number. No rule for such cases can be given, except that the pupils should be directed to ask the question: To what unit is the measure correct? If the zero is located in the place of significance of a unit in terms of which a measure is stated, then it is significant, otherw[illegible] not significant. For example, if the distance from the eart[illegible] is 240,000 miles, correct to the nearest mile, then [illegible]t number are significant. If this same distance i[illegible]o the nearest thousand miles, then the first zero [illegible] in that number is significant, and the last three zeros [illegible] significant. A very effective method of writing such numbers is as follow[illegible]keep the zeros that are significant and we write the number [illegible] *form.* Thus the above distance from the earth to the moon would be written as $2.40 \cdot 10^5$. If there were only two significant digits in the above number, it would be written as $2.4 \cdot 10^5$.

D. THE TERMINATION OF THE SIGNIFICANT DIGITS ON THE RIGHT OF AN APPROXIMATE NUMBER IN THE PLACE OF SIGNIFICANCE WHERE THE SIGNIFICANT DIGIT ON THE LEFT OF THE APPARENT ERROR IS LOCATED

I.P.M. Examples of approximate numbers whose apparent errors are given showing that the digits, if any, located in the place of significance where the first significant digit on the left of the apparent error is, as well as all the digits to the right of this place of significance, are not significant. For example, in $67{,}532 + 50$ the digits 3 and 2 are not significant.

T.R. The importance of this topic cannot be overestimated. The analysis of the approximativeness of the numerical results in the work in elementary algebra and numerical trigonometry is effected by means of the apparent errors only. The teacher should stress the fact that in the numerical value of a measure, or, generally, in an approximate number, the digits in the place of significance, where the first significant digit on the left of the apparent error appears, and all the digits to the right of this place of significance are of no importance so far as the statement of the magnitude of a measurement is concerned. In other words, even if there were some information concerning the magnitude of those units and of the units appearing to the right, little use could be made of that information, except for the information regarding the error. Since there is an error in every measurement, and this error is expressed in terms of a certain unit, this unit and the digits appearing in the place of significance of this unit are not significant in the statement of that measurement. Thus the apparent error of an approximate number serves the purpose of denoting the place of significance where the appearance of significant digits of an approximate number terminates.

IV. ROUNDING OFF DIGITS (Appendix 1)

The purpose of this section is to develop the technique of rounding off digits and at the same time to show the connection between the processes of rounding off digits and the processes of measurement. The rules for rounding off digits should be shown as being not arbitrary rules, but agreements of practice common in all fields of human activity, wherever measurements are used, and based on the nature of the processes of measurement.

A. THE RULES FOR ROUNDING OFF DIGITS

I.P.M. Examples showing the unimportance of the non-significant digits so that they should be disposed of as facts irrelevant to the information concerning measures. For example, in the number 34.256 ± 0.5 all the digits to the right of the decimal point are useless.

T.R. The pupils should be shown that the process of rounding off digits is a necessary and important procedure in Approximate Computation. Rounding off digits is necessary for the proper statement of a measure or of a numerical result in a manner and form consistent with their respective precisions as determined by the apparent errors. The pupils should become convinced that a numerical result must not contain irrelevant information usually given by non-significant digits.

i. The forcing of a digit

I.P.M. Examples showing that rounding off digits should follow the method of stating a measure to the nearest unit. That is, if the first digit to be dropped is 5, 6, 7, 8, or 9, then the nearest unit is one greater than the one represented by the first significant digit on the extreme right of a number that is to be retained. For example, the number 47.435 ± 0.005 is rounded off to 47.44; the number 62.36 ± 0.05 is rounded off to 62.4. If the number is integral then the dropped digits are replaced by non-significant zeros.

T.R. (See Appendix 1, Rule 3.) The pupils are already familiar with the practice of the rule described in this case. At this stage it is only necessary to review some examples in measurement and generalize this practice into a rule. The only difficulty that may arise in this case is in connection with the digit 5. Some writers use a rule which suggests the forcing of an odd significant digit only, when the first digit to be dropped is 5; and if the first digit to be dropped is 5 and the first significant digit on the extreme right of a number that is to be retained is even, then it should be left unchanged. This rule is entirely arbitrary, and there is little justification for basing its use on the theory of probability of the compensation of errors, as some writers state. From the pedagogical point of view this rule complicates matters instead of simplifying the general rule. If there is any justification (from the purely theoretical point of view) for its use it has yet to be demonstrated. Moreover, it is just as logical and just as consistent to force an even significant digit and to leave an odd significant digit unchanged. It is better to disregard the case of the digit 5 completely and to treat this digit in the same manner as all the digits greater than 5.

ii. The dropping of a digit

I.P.M. Further examples showing that if the first digit to be dropped is 0, 1, 2, 3, 4, or 5, then the first significant digit on the extreme right of the number to be retained should be left unchanged, while the non-significant digits are dropped. If the number is integral then the dropped digits are replaced by non-significant zeros. For example, the number 34.256 ± 0.5 is rounded off to 34.; the number 34,256 ± 500 is rounded off to 34,000.

T.R. This unit represents a continuation of the development of the rule for rounding off digits. An important procedure in the process of rounding off digits is the replacement of the dropped digits of an integral number by zeros. This is done because the dropped digits are of no value as far as information regarding the measure is concerned and the proper digits to be used instead of the dropped digits are the non-significant zeros which denote "nothing."

iii. The writing of the non-significant zeros (optional)

I.P.M. The writing of the non-significant zeros in smaller script as follows: 34,000.

T.R. Some European writers suggest the writing of the non-significant zeros as well as the non-significant digits in a script smaller than the script of the significant digits. This practice is not yet universally accepted in the field of mathematics. This method of writing would enable one to distinguish between the zeros which are significant and the zeros which are not significant. On the other hand, such a practice might prove to be difficult to follow in secondary school mathematics because it would place extra responsibility on the pupils to keep track of the two kinds of zeros. This possibility, however, should not preclude the introduction of this method of writing the zeros in secondary school mathematics.

B. THE EFFECT OF ROUNDING OFF DIGITS ON THE PRECISION OF A MEASUREMENT

I.P.M. Analysis of rounded off numbers showing that a rounded off number represents a measure which is less precise. For example, the number 34.573 represents a measure correct to thousandths of a unit, whereas the number 34.6 represents a measure correct only to tenths of a unit.

T.R. The pupils should learn to recognize the fact that rounding off digits lowers the precision of a measure whose value is represented by an approximate number. The teacher should use ample illustrative materials for the purpose of developing the technique of rounding off digits as well as for the purpose of elucidating the idea of the dependence of precision on rounding off digits.

V. The Relative Error (pages 125–131)

The purpose of this section is to develop the concept of the relative error and its properties as well as to show its importance in Approximate Computation. The relative error represents a certain measure of approximativeness. Moreover, the comparison of two measures is more often effected by means of the relative error than by means of the apparent error. The pupils should see that whereas the apparent error analysis concerns itself with the positional location of the last significant digit on the extreme right of an approximate number, the relative error concerns itself with the entire approximate number. This comparison of the two types of analysis should not lead, however, to the impression that the relative error analysis is always more powerful than the apparent error analysis. These two methods are equally important in the processes of Approximate Computation in general, but in some cases it is advantageous to use apparent error and in some cases it is advantageous to use relative error. In all cases of Approximate Computation both types of error are important and their combined use is absolutely essential. For the purposes of elementary algebra and numerical trigonometry, however, we shall make use of the apparent error almost exclusively.

A. THE PROPERTIES OF THE RELATIVE ERROR

T.R. The pupils should learn that the relative error is not obtained directly from the process of measurement. Moreover, the pupils should understand that neither the apparent nor the relative errors are exact numbers; they are both approximate. This fact, however, does not deter us from using them. Under the circumstances we must be content to make the best use of whatever information we can secure under the most unfavorable conditions. The pupils should also make a practice of writing the relative errors with one or at most two significant digits only. For the purposes of elementary algebra and numerical trigonometry this mode of writing is quite appropriate, and it should not cause any difficulties from the point of view of the arithmetical work involved in it.

i. The ratio property

I.P.M. The definition of a relative error as $r = \frac{e}{A}$, where e is the apparent error, and A is the approximate number. Numerical problems leading to calculation of the relative errors.

T.R. The pupils should be given ample opportunity to perform numerical work in obtaining relative errors which should be expressed as decimal fractions or as percentage errors. The pupils should be led to see that since the relative error is a ratio of the apparent error to the approximate number (or the ratio of the error in measurement to the numerical value of the measure), it is a measure of relationship between these two magnitudes (see sec. *ii* below).

ii. The amount of error per unit of measure in terms of which the measure is expressed. The percentage error

I.P.M. Numerical problems showing this property of relative errors of approximate numbers. Practice in expressing the amount of error per unit of measure in terms of per cents.

T.R. Since the relative error is a ratio of the apparent error to the approximate number, it expresses the measure of the amount of error per unit of measure in terms of which the measure is expressed. This amount is expressed in terms either of a decimal fraction or of a percentage.

iii. The concept of accuracy of a measure

I.P.M. Numerical problems leading to the comparison of two measures in order to determine the more accurate one. For example, 34.1 in. and 1,256 ft. The two relative errors are:

$$\frac{0.05}{34.1} \doteq 0.0015 \qquad \frac{0.5}{1{,}256} \doteq 0.0004,$$

(the symbol $\doteq$ denotes approximate equality) or 0.15% and 0.04% respectively. The second measure is decidedly the more accurate one, whereas the first measure is the more precise one.

T.R. Comparison of the approximativeness of two measures is made by means of their respective errors. In the same way that the precisions of two measures are compared, we establish a comparison in terms of a new concept, that is, of *accuracy*. The measure whose relative error is smaller, that is, the measurement having a smaller error per unit of measure in terms of which it is expressed, is the more accurate.

iv. The relation of the relative error to significant digits

I.P.M. Examples similar to those in sec. *iii* above showing that the smaller the relative error the greater is the number of significant digits that can be expected in an approximate number.

T.R. The relationship between relative errors and the number of significant digits of their respective approximate numbers should be stressed from the point of view of conveying the idea that the more significant digits there are in an approximate number the more accurate is the measure whose value is represented by that approximate number. This concept is very important in Approximate Computation in view of the fact that it will enable the pupils to understand later on why they should not expect to obtain more significant digits in their numerical results than there are significant digits in the least accurate of the numbers with which they work. (The only exceptions are some cases of root extraction.)

B. THE RELATION BETWEEN THE APPARENT AND THE RELATIVE ERRORS AND THE APPROXIMATE NUMBER

I.P.M. The use of the formula $e = rA$ for the determination of the apparent error of an approximate number. For example, the approximate number is 3,247, and it has a relative error of 0.05% (or simply an error of 0.05%). It is required to determine the correct number of significant digits in this approximate number. The relative error is 0.0005. The apparent error is determined as follows:

$$0.0005 \cdot 3{,}200 = 1.6.$$

Rounding off, we get an apparent error $e = 2$. The approximate number is then 3,250, correct to three significant digits.

C. THE EFFECT OF ROUNDING OFF DIGITS ON THE ACCURACY OF A MEASURE

I.P.M. Analysis of rounded off numbers showing that a rounded off number represents a measure which is less accurate. For example, the number 34.573 has a relative error:

$$\frac{0.0005}{34.573} = 0.000014,$$

while the number 34.6 has a relative error:

$$\frac{0.05}{34.6} = 0.0014.$$

T.R. Rounding off the digits of an approximate number affects the accuracy of the numerical value of the measure in the same manner and the same direction as it affects the precision. A rounded off approximate number represents a less accurate measure, and this fact should be illustrated with numerical examples. The pupils should work out many examples in order that they may become convinced of the relationship between accuracy and the number of significant digits of an approximate number. This property is very important in the processes of Approximate Computation, and much use will be made of it in all future work.

VI. The Relation Between the Magnitudes of Approximate Numbers and the Magnitudes of Their Respective Apparent Errors

I.P.M. Numerical examples for the purpose of comparison of the magnitudes of two measures; that is, the approximate number and its apparent error.

Numerical examples in multiplication of two approximate numbers such as: (34.5 ± 0.05)(423. ± 0.5), or (3,450 ± 5)(4,200 ± 50), showing that the products $0.05 \cdot 0.5 = 0.025$ and $5 \cdot 50 = 250$ are so small in comparison with the products of their respective approximate numbers that they can be disregarded.

T.R. The pupils should learn that in comparison with the magnitude of an approximate number the magnitude of its apparent error is considered to be very small. The teacher should explain that in order to be able to judge whether a thing is very small or not, one must have another object with which this thing should be compared. A resourceful teacher will find many illustrations for classroom presentation and for elucidation of the idea of relativity of magnitude. Many interesting illustrations from the physical world with their appropriate numerical values will not only prove useful, but will, no doubt, fascinate the pupils. For example, the comparison of the numerical values of the diameters of the earth and the sun, the comparison of the distances from the earth to the moon and to the sun, the comparison of the distances from the earth to the sun and to the nearest star, or the comparison of the weight of an airplane and of a fly.

The purpose of this section is to develop the necessary background for the derivation of the formulas for errors of all numerical processes involving multiplication in one form or another. The pupils should convince themselves that when errors are multiplied by errors the products represent quantities so small in comparison with the magnitudes of the products of their respective approximate numbers that the magnitudes of the products of the errors produce no effect on the precision or accuracy of the final numerical results of the performed numerical operations with approximate numbers. It is, therefore, quite appropriate and permissible to disregard and to discard the products of errors by errors.

CHAPTER VI

APPROXIMATE COMPUTATION: NUMERICAL OPERATIONS

THE rules for Approximate Computation suitable for the level of elementary algebra and numerical trigonometry developed below assume that all approximate numbers have apparent errors equal to ±0.5 of a unit of measure in terms of which the measures (or their numerical values) are expressed. The evaluation of the apparent errors of the numerical results of the performed computations should have as its aim the determination of the place of significance of the last significant digit on the extreme right of an approximate number. Strictly speaking, this does not represent a perfectly true account of the approximativeness of the numerical result, because no other definite use of the apparent errors is made, and their numerical value is not taken into consideration in the statement of the final numerical result of the performed computations. The first step in Approximate Computation, however, should be concerned with significant digits only, and the use of apparent errors should be restricted to the study of their influence on the results of computations with approximate numbers; that is, to the determination of the place of significance of an approximate number where the error appears, and hence where no reliability could be placed, regardless of the magnitude of the error.

The attention of the teacher is called to the so-called numerical exercises in geometry in which Approximate Computation is very important. Numerical exercises in geometry should involve data representing measurements. This is directly in keeping with the ideas of the metric properties of Euclidean geometry. But in some cases, where the illustration of these metric properties is more important than the actual numerical result, some geometrical problems may be considered "theoretical," and in such cases the numerical data concerning the magnitude of the geometrical figure (or of its elements) may be considered exact. However, the teacher should be always aware of the fact that even in the case of "theoretical" problems many numerical results may not be exact. Such cases arise when the expressions to be evaluated involve irrationalities or lead to repeating and (or) non-terminating decimals. The teacher should avoid using theoretical problems except for illustration purposes.

The development of the rules for Approximate Computation below will be made algebraically. The following symbolism is quite appropriate for these purposes:

$a, b, c, \ldots$ will denote approximate numbers.
$A, B, C, \ldots$ will denote their true value portions.
$e_1, e_2, e_3, \ldots$ will denote their apparent error.

Then

$$a = A + e_1,$$
$$b = B + e_2,$$
$$c = C + e_3,$$
$$\cdots\cdots\cdots\cdots$$

The following very important principle should be enunciated by the teacher as early as possible. In Approximate Computation all approximate numbers that are involved in a certain numerical process (or several numerical processes) should preferably and if possible contribute equal shares of error, unless there is a good reason for allowing a given approximate number (in the case of multiplication or division) to contribute a smaller error. The kind of error is not specified in the above statement because in addition or subtraction this principle refers to apparent errors, and in all other numerical processes it refers to relative errors. The meaning of this statement will become more clear in the analysis of the respective processes.

1. ADDITION (pages 138–142)

It is important to bear in mind that in the case of addition the criterion of precision in judging approximate numbers is the principle on which the rule for addition is based. It is important that the approximate numbers to be added represent measurements of the same precision because the sum of two measures of different precisions cannot be as precise as the more precise measure of the two. For example, 15.26 and 3.4725, representing two measures of which the first is the less precise one, illustrate this idea. Their sum, 18.7325, would be misleading because in obtaining this particular sum we made an assumption that the first measure was of the form 15.2600. Actually, this assumption cannot be justified at all because the measure represented by the number 15.26 may be any measure lying between 15.255 and 15.265. It follows, then, that the number 3.4725 contains more information than is needed, and some of it should be discarded as irrelevant. In other words, it should be rounded off to the same units of measure in terms of which the number 15.26 is expressed.

A. THE INFLUENCE OF THE APPARENT ERRORS ON THE SUM OF TWO APPROXIMATE NUMBERS

I.P.M. The sum of two approximate numbers

$$\begin{aligned} a &= A + e_1 \\ b &= B + e_2 \end{aligned}$$

is

$$a + b = A + B + (e_1 + e_2)$$

The apparent error of the sum of two approximate numbers may numerically be equal to the sum of the respective apparent errors of the two numbers. For example,

$$\begin{array}{r} 6.25 \pm 0.005 \\ 13.38 \pm 0.005 \\ \hline 19.63 \pm 0.01 \end{array}$$

The sum is then 19.6. Also

$$\begin{array}{r} 17.98 \pm 0.005 \\ 32.87 \pm 0.005 \\ \hline 50.85 \pm 0.01 \end{array}$$

The sum is then 50.9.

T.R. The derivation of the expression for the apparent error of the sum of two approximate numbers should lead the pupils to the realization of the fact that the apparent errors of the individual approximate numbers influence the sum of these numbers in such a manner that the last significant digit on the extreme right of the resultant sum ceases to be significant. This is because each of the respective apparent errors is ±0.5 of a unit in terms of which the measurement is expressed.

B. THE INFLUENCE OF THE APPARENT ERRORS ON THE SUM OF MORE THAN TWO APPROXIMATE NUMBERS

I.P.M. The sum of three approximate numbers

$$\begin{aligned} a &= A + e_1 \\ b &= B + e_2 \\ c &= C + e_3 \end{aligned}$$

is

$$a + b + c = A + B + C + (e_1 + e_2 + e_3)$$

The apparent error of the sum of three approximate numbers may numerically be equal to the sum of the respective apparent errors of the three numbers. For example,

$$\begin{array}{r} 27.43 \pm 0.005 \\ 125.62 \pm 0.005 \\ 37.32 \pm 0.005 \\ \hline 190.37 \pm 0.015 \end{array}$$

The sum is then 190.4.

A discussion leading to the conclusion that in the case of a sum of 20 approximate numbers, two digits on the extreme right of the sum cease to be significant.

T.R. This section should be devoted to the generalization of the expression for the apparent error of the sum of several approximate numbers. The teacher should use many numerical examples leading to the generalization of this rule, especially examples illuminating the fact that the loss of one significant digit on the extreme right of a sum is actually a very slow process. Whereas in the case of the sum of two approximate numbers one digit is lost, it would take 20 approximate numbers to cause a loss of two digits, and it would take 200 approximate numbers to cause a loss of three significant digits, in the respective sums.

C. THE RULE FOR ADDITION OF APPROXIMATE NUMBERS

I.P.M. Examples in obtaining sums of approximate numbers without the use of apparent errors according to the rule that:

In the sum of two approximate numbers one significant digit on the right is lost.

In the sum of 20 (up to, but not including, 200) approximate numbers, two significant digits on the right are lost.

When two approximate numbers are given that do not represent measures of the same precision, the more precise approximate number should be rounded off so that the two approximate numbers to be added will be of the same precision, that is, of the same precision as the number representing the less precise measure. For example, 51.273 and 62.4 are to be added. The sum is obtained as follows:

$$\begin{array}{r} 51.3 \\ 62.4 \\ \hline 113.7 \end{array}$$

and the sum, correctly stated, is 114.

If more than two approximate numbers are given, and not all of them are of the same precision, then in order to prevent unduly large apparent errors of sums in such cases, the sum of these approximate numbers is obtained in several parts. We start with the more precise approximate numbers, and we obtain a partial sum according to the rule above. This partial sum is rounded off to conform with the precision of the next less precise number, or numbers, and a partial sum is obtained. We proceed in this manner again. The newly obtained partial sum is rounded off to conform with the precision of the next less precise number, or numbers, and a new sum is obtained. This process is repeated again, until step by step, by the method of exhaustion, we reach the stage at which the total sum may be obtained. For example, the following approximate numbers are to be added: 24.8364, 342.2, 67.453, 873.4523, 92.34283. We proceed then as follows:

	92.3428
	873.4523
	24.8364
	990.6315
Then	990.632
	67.453
	1,058.085
Finally	1,058.1
	342.2
	1,400.3

and the correctly stated sum is 1,400.

T.R. The formulation of the statement of the rule for addition of approximate numbers should be followed up with many numerical examples so that a correct technique and a habit of working with approximate numbers are developed. It is important that the pupils appreciate the necessity of stating numerical results in terms of significant digits only. The process of developing this habit may be at first slow, but once the notion of the approximativeness of numerical results is developed, further work with approximate numbers will, no doubt, be a matter of common procedure. The case of a sum of several approximate numbers representing measures of different precision should be given particular attention because this case will impress the pupils with the importance of following proper procedure when working with approximate data, especially with regard to the necessity of trying to retain a maximum approximativeness under adverse conditions which generally govern all cases of Approximate Computation. The teacher should use many numerical examples to illustrate the fact that unduly rash rounding off may very often lead to numerical results that are far from being correct. For example, the sum of the following numbers may represent a typical case: 127.6, 625.27, and 343.56 are to be added, and each number is approximate. If we round off the last two numbers and perform the addition we have

$$\begin{array}{r} 127.6 \\ 625.3 \\ 343.6 \\ 1{,}096.5 \end{array}$$

and the final result is 1,097. But if we add the last two numbers first, and round them off, and then add the first number, we get

$$\begin{array}{r} 625.27 \\ 343.56 \\ \hline 968.83 \\ \hline 968.8 \\ 127.6 \\ \hline 1{,}096.4 \end{array}$$

and the final result is 1,096., which is closer to the true value.

II. SUBTRACTION (pages 142–145)

The development of the rule for subtraction of approximate numbers should follow the same steps as in addition. Some difficulty may present itself in the case of the derivation of the expression for the apparent error of the difference of two approximate numbers. The teacher should call the pupils' attention to the fact that the apparent error of a difference of two approximate numbers, whose respective apparent errors both lie in either direction, that is, are ±0.5 [illegible] a unit of measure in terms of which the measure is expressed, may be equal numerically to the sum of these errors.

All further remarks concerning subtraction are identical in nature and content with the remarks concerning addition, and the reader is referred to section I above.

A. THE INFLUENCE OF APPARENT ERRORS ON THE DIFFERENCE OF TWO APPROXIMATE NUMBERS

I.P.M. The difference of two approximate numbers

$$\begin{array}{lr} & a = A + e_1 \\ & b = B + e_2 \\ \cline{2-2} \text{is} & a - b = A - B + (e_1 + e_2) \end{array}$$

The apparent error of the difference of two approximate numbers may numerically be equal to the sum of the respective apparent errors of the two numbers. For example,

$$\begin{array}{r} 423.5 \pm 0.05 \\ 221.6 \pm 0.05 \\ \hline 201.9 \pm 0.1 \end{array}$$

The difference is 202.

Numerical examples illustrating the fact that the relative error of the difference of two approximate numbers is much greater than the relative error of the minuend. In other words, the difference of two approximate numbers is always less accurate than the minuend.

B. THE RULE FOR SUBTRACTION OF APPROXIMATE NUMBERS

I.P.M. Examples in obtaining the difference of two approximate numbers without the use of apparent error according to the rule that in the difference of two approximate numbers one digit on the extreme right is lost.

When one of the approximate numbers given represents a measure of greater precision than the other, then the approximate number representing this measure should be rounded off so that it is of the same precision as the number representing the measure of lesser precision. For example, if 151.273 is the minuend and 62.4 is the subtrahend, then the subtraction is performed as follows:

$$\begin{array}{r} 151.3 \\ 62.4 \\ \hline 88.9 \end{array}$$

and the difference is 89.

III. Multiplication (pages 145–164)

In the processes of multiplication, division, raising to a power, and extraction of roots the use of relative errors for the purpose of evaluating the approximativeness of numerical results is more effective than the use of apparent errors. This, however, refers to the general theoretical approach to these processes. In elementary algebra and numerical trigonometry the pupils are expected to use apparent errors mostly. But the analysis of these processes even on the levels just stated will involve the relative errors to the extent that, without making the presentation unduly difficult, these errors will enable the teacher to present a clear, concise, and convincing analysis of the respective operations. It should be the aim of the teacher to direct the pupils' reasoning in order to show that the analysis of the operations is effected as rigorously as any other process they study in connection with Approximate Computation. The pupils should appreciate the fact that the work with approximate numbers—and the above stated operations serve this purpose best—besides being a purely mathematical process, is a very important tool useful in many fields of human endeavor where qualitative and quantitative properties of objects are expressed in terms of numbers.

The development of the rules for multiplication of approximate numbers is a somewhat lengthy process. It is best for the pupils to go through this development slowly. If sufficient time is spent on the analysis of multiplication, there will be no difficulty with the analysis of division, raising to a power, and the extraction of roots. The teacher should remember that in multiplication as well as in the three other processes the pupils will have to make exclusive use of significant digits. The teacher should not fail to recall to the pupils the close relationship between the magnitude of the relative errors and the number of significant digits of approximate numbers.

A very important rule should be emphasized by the teacher throughout the entire work that follows. This rule is that all numerical results, before they are stated in final form, should be obtained with at least one more digit than the number of significant digits allowed by the approximate data. Then this last digit is rounded off. The importance of this rule is obvious, and ample numerical illustrations will enable the teacher to demonstrate this fact to the pupils.

A. THE FORMULA FOR THE APPARENT ERROR OF A PRODUCT OF TWO APPROXIMATE NUMBERS

I.P.M. The product of two approximate numbers

$$\begin{aligned} a &= \text{[illegible]} + e_1 \\ b &= \text{[illegible]} + e_2 \end{aligned}$$

is

$$ab = AB + (Ae_2 + Be_1) + e_1e_2$$

Disregarding e_1e_2, we have the expression for the apparent error of the product: $Ae_2 + Be_1$. For example, let $a = 47.5 \pm 0.05$, and $b = 27. \pm 0.5$. Then $ab = 1{,}282.5 \pm (23.75 + 1.35) = 1{,}282.5 \pm 25.10$. After rounding off, the product is 1,300, correct to two significant digits.

T.R. The derivation of the formula for the apparent error of the product of two approximate numbers should make use of a result already obtained in Chapter V, sec. VI. There it was found that the product of two apparent errors is so small in comparison with the product of their respective approximate numbers that there is justifiable reason for disregarding the product e_1e_2. Numerical examples should illustrate this fact fully and convincingly. The pupils will be expected to make use of the expression for the apparent error of a product of two approximate numbers, and therefore they should be given a sufficient number of problems for them to become fully conversant with this process.

B. THE RELATIVE ERROR OF A PRODUCT OF TWO APPROXIMATE NUMBERS

I.P.M. The relative error of the product of two approximate numbers is

$$\frac{Ae_2 + Be_1}{AB} = \frac{e_1}{A} + \frac{e_2}{B}.$$

That is, it may be equal to the sum of the relative errors of the respective approximate numbers.

Numerical examples leading to the calculation of the relative errors of products of two approximate numbers. For example, let $a = 47.5 \pm 0.05$, and $b = 27. \pm 0.5$. The relative errors are:

$$\frac{0.05}{47.5} \quad \text{and} \quad \frac{0.5}{27.}$$

The relative error of the product is

$$\frac{0.05}{47.5} + \frac{0.5}{27.} = 0.00105 + 0.0185 = 0.02.$$

This result checks with the relative error of the product obtained by means of the apparent error of the product:

$$\frac{25.1}{1{,}282.5} = 0.0195 = 0.02.$$

T.R. The derivation of the expression for the relative error of the product of two approximate numbers should have as its aim the demonstration of the fact that the product of two approximate numbers cannot be as accurate as any of the respective factors. This fact should be illustrated with numerical examples. The purpose of this section is to lead up to the development of the rule for multiplication in terms of significant digits.

C. THE SIGNIFICANT DIGITS OF THE PRODUCT OF TWO APPROXIMATE NUMBERS

I.P.M. Numerical examples illustrating the fact that the product of two approximate numbers, both given as correct to k significant digits each, may be correct to k or to $(k - 1)$ significant digits. For example, $a = 42.3 \pm 0.05$, and $b = 12.2 \pm 0.05$. Then

$ab = 516.06 \pm 0.05(42.3 + 12.2) = 516.06 \pm 2.725$, and $ab = 520$.

Also $a = 42.3 \pm 0.05$, and $b = 32.2 \pm 0.05$. Then

$ab = 1{,}362.06 \pm 0.05(42.3 + 32.2) = 1{,}362.06 \pm 3.725$, and $ab = 1{,}360$.

Numerical examples illustrating the fact that the product of two approximate numbers, one given as correct to k significant digits, and the other as correct to $(k + r)$, $r = 1, 2, 3, \ldots$, significant digits is correct to k or $(k - 1)$ significant digits. For example, $a = 576.43 \pm 0.005$, and $b = 1.22 \pm 0.005$. Then

$ab = 703.2446 \pm 0.005(576.43 + 1.22) = 703.2446 \pm 1.88825$, and $ab = 700$.

Also $a = 576.43 \pm 0.005$, and $b = 2.22 \pm 0.005$. Then

$ab = 1{,}278.6746 \pm 0.005(576.43 + 2.22) = 1{,}278.6746 \pm 2.89325$, and $ab = 1{,}280$.

T.R. The theoretical development of this topic is much too difficult for the level of elementary algebra and numerical trigonometry. However, continuing the discussion of sections A and B above, the teacher should use many examples showing the two following cases that may arise in the process of multiplication of two approximate numbers, when both approximate numbers are given as correct to the same number of significant digits, for example, k each. If the product of the first significant digits on the extreme left of the approximate numbers is equal to or is greater than 10, or if the product of the first two significant digits of each approximate number is equal to or is greater than 1,000, then the product of the two approximate numbers is correct to k significant digits. Otherwise, the product is correct to $(k - 1)$ significant digits. In a similar manner the teacher should illustrate with numerical examples the fact that the product of two approximate numbers, one correct to k significant digits, and the other correct to $(k + r)$, $r = 1, 2, 3, \ldots$, significant digits, is correct to k or $(k - 1)$ significant digits, and that the above test of the product of the first significant digits on the left of the approximate numbers is also applicable in this case. The numerical work should precede the drawing of any conclusions, and it should be based on the calculation of the apparent error of the product, and thus determine the place of significance in the product where the error occurs. The teacher should illustrate this topic with ample numerical examples so that the facts just stated will be fully demonstrated. The statement of the rules for multiplication of two approximate numbers depends entirely on the development of the material in the three sections above; this development should be very thorough.

D. THE RULES FOR MULTIPLICATION OF TWO APPROXIMATE NUMBERS

I.P.M. The statement and illustration of the rules for multiplication of two approximate numbers in terms of significant digits.

i) If two approximate numbers are given such that one is correct to k significant digits, and the other is correct to $(k + r)$, $r = 0$, 1, 2, 3, . . . , significant digits, then if the product of the first significant digits on the extreme left of the approximate numbers is equal to or is greater than 10, or if the product of the first two significant digits on the extreme left of each of the approximate numbers is equal to or is greater than 1,000, then the product of the two approximate numbers is correct to k significant digits. Otherwise, the product is correct to $(k - 1)$ significant digits.

ii) If one of the approximate numbers is correct to k significant digits, and the other approximate number is correct to $(k + r)$ significant digits $(r = 2, 3, 4, \ldots)$, then the second approximate number should be rounded off to $(k + 1)$ significant digits.

T.R. For the purposes of elementary algebra and numerical trigonometry the rules for multiplication should be stated in terms of significant digits. The pupils should be given to understand the reason for keeping one extra digit in one of the factors, when such a procedure is possible. The use of one extra digit in one of the factors results in a smaller relative error of this factor, and thus we may obtain a smaller apparent error of the product. It should be pointed out that later on, in intermediate algebra, advanced algebra, and trigonometry, this procedure will be used very often. It should be also pointed out to the pupils that there is no need of using more than one extra significant digit. The use of two or more extra significant digits would not result in a product that approximates any more closely, and on the other hand, it would entail additional and unnecessary numerical work. Thus, $a = 18.34$, $b = 4.5826$, and $ab = 84.0$ (error of product is 0.025); $a = 5.834$, $b = 3.141592$, and $ab = 18.33$ (error of product is 0.002). The pupils should develop the habit of rounding off approximate numbers in accordance with the rules stated above, before the actual process of multiplication takes place.

E. THE DETERMINATION OF THE NUMBER OF SIGNIFICANT DIGITS OF FACTORS REPRESENTED BY MATHEMATICAL CONSTANTS SUCH AS π, $\sqrt{2}$, $\sqrt{3}$, and $\sqrt{5}$

I.P.M. Numerical illustrations showing that certain numerical constants, such as π, $\sqrt{2}$, $\sqrt{3}$, and $\sqrt{5}$, may be determined to any number of significant digits, and that these constants should be used in accordance with the rules stated in section D above.

T.R. The subject that is discussed in this section is very important because considerable use will be made of this procedure in the processes of numerical trigonometry as well as in any other case when opportunity may present itself. The pupils should be led to realize that by means of using an extra significant digit they may reduce the amount of error in the final numerical result of the performed operations. Even though the actual advantage of this procedure may not be obvious in the work in elementary algebra, because at this level the statement of the error of the final numerical result does not enter into consideration, the pupils should form a habit of using an extra significant digit in order that they may learn this procedure and carry over this rule into their later work in mathematics.

F. THE PRODUCT OF AN EXACT NUMBER AND AN APPROXIMATE NUMBER

I.P.M. Illustrative examples showing that in case one of the factors is represented by an exact number and the second factor is approximate, the same rules hold; that is, the apparent error of the exact number is zero, and the apparent error of the product is obtained in the same manner.

T.R. Often one factor may be exact, and the pupils should learn how to work with a combination of exact and approximate numbers. The nature of exact numbers is such that their errors, both apparent and relative, are zero.

G. THE PRODUCT OF THREE APPROXIMATE NUMBERS

I.P.M. The product of three approximate numbers

$$a = A + e_1$$
$$b = B + e_2$$
$$c = C + e_3$$

is $abc = ABC + ABe_3 + ACe_2 + BCe_1 + Ae_2e_3 + Be_1e_3 + Ce_1e_2 + e_1e_2e_3,$

and the apparent error of the product is

$$ABe_3 + ACe_2 + BCe_1.$$

The relative error of the product is

$$\frac{ABe_3 + ACe_2 + BCe_1}{ABC} = \frac{e_1}{A} + \frac{e_2}{B} + \frac{e_3}{C}.$$

Numerical illustrations showing the advantage in using one or two factors with an extra significant digit.

T.R. The teacher should not fail to present this topic which is very important in numerical trigonometry and geometry. Multiplication of three binomials, which is necessary for the derivation of the expression of the apparent error of the product of three approximate numbers, will not present much difficulty if the pupils know how to multiply a polynomial by a binomial. It is perhaps tedious to obtain the numerical value of the apparent error of a product of three approximate numbers by means of the formula for the apparent error. For example, obtain the volume of a box whose dimensions are $l = 4.57$ in., $w = 3.37$ in., and $h = 3.18$ in. Answer: 49. in^3 (the error is 0.3). The teacher may revert to the method of securing the apparent error of this product by means of the relative error of the product and the first few significant digits on the left of this product (see pages 125–130). All the rules concerning the number of significant digits obey the same principles stated in D above, and the teacher should make use of numerical illustrations for this purpose.

IV. DIVISION (pages 170–180)

The derivation of rules for division of approximate numbers in terms of significant digits is considerably more difficult than the derivation of the rules for multiplication. As will be shown below, the relative error of a quotient of two approximate numbers is identical in form with the relative error of a product of two approximate numbers. In order to avoid difficult discussions the teacher may direct the pupils' attention to this fact and base the statement of the rules for division on the reasonableness of the expectation that the rules that are about to be stated are similar in nature and content to the rules for multiplication.

All remarks made in regard to multiplication of two approximate numbers are applicable to the case of division, and the reader is referred to section III, A, B, C, and D above.

A. THE FORMULA FOR THE APPARENT ERROR OF A QUOTIENT OF TWO APPROXIMATE NUMBERS

I.P.M. Let the two approximate numbers be

$$a = A + e_1$$

and

$$b = B + e_2$$

Take the product $ab = AB + (Ae_2 + Be_1)$ and divide the left side of the equality by b^2 and the right side of the equality by B^2. This will destroy the equality because B^2 is not equal to b^2, but this change is not materially large enough to concern us. We have then

$$\frac{a}{b} = \frac{A}{B} + \frac{Ae_2 + Be_1}{B^2},$$

and the apparent error of the quotient $\frac{a}{b}$ is

$$\frac{Ae_2 + Be_1}{B^2}.$$

Numerical examples and problems in obtaining a quotient of two approximate numbers and the apparent error of this quotient. For example, $a = 42.3 \pm 0.05$, $b = 12.2 \pm 0.05$. Then

$$\frac{a}{b} = \frac{42.3}{12.2} + \frac{0.05(42.3 + 12.2)}{(12.2)^2} = 3.47 \pm 0.02.$$

The quotient is then $3.\bar{5}$.

T.R. Another method of deriving the expression for the apparent error of a quotient of two approximate numbers is as follows: We form the maximum and minimum quotients:

$$\text{minimum: } \frac{A - e_1}{B + e_2}, \qquad \text{and maximum: } \frac{A + e_1}{B - e_2}.$$

Half the difference of these two quotients is numerically equal to the apparent error of the quotient, that is:

$$\tfrac{1}{2}\left[\frac{A+e_1}{B-e_2}-\frac{A-e_1}{B+e_2}\right]=\frac{(A+e_1)(B-e_2)-(A-e_1)(B+e_2)}{2(B^2-e_2^2)}$$

$$=\frac{Ae_2+Be_1}{B^2}.$$

This method of derivation, however, implies some notions from the theory of inequalities, which form a part of intermediate algebra.

The teacher should point out to the pupils that in the case of division the proper procedure is to obtain the apparent error of the quotient first, and then proceed with the division, terminating in the place of significance or in one place beyond the place of significance where the error occurs.

B. THE RELATIVE ERROR OF A QUOTIENT OF TWO APPROXIMATE NUMBERS

I.P.M. The relative error of the quotient of two approximate numbers is

$$\frac{Ae_2+Be_1}{B^2}:\frac{A}{B}=\frac{e_1}{A}+\frac{e_2}{B},$$

that is, it may be equal to the sum of the relative errors of the two respective approximate numbers.

Numerical problems leading to the calculation of the relative error of a quotient of two approximate numbers. For example, let $a = 42.3 \pm 0.05$, and $b = 12.2 \pm 0.05$. The relative errors are

$$\frac{0.05}{42.3} \quad \text{and} \quad \frac{0.05}{12.2}.$$

The relative error of the quotient is

$$\frac{0.05}{42.3}+\frac{0.05}{12.2}=0.00118+0.00410=0.0053.$$

This result checks with the relative error of the quotient obtained by means of the apparent error of the quotient:

$$\frac{0.05(42.3+12.2)}{148.84}:3.47=0.0053.$$

C. THE SIGNIFICANT DIGITS OF A QUOTIENT OF TWO APPROXIMATE NUMBERS

I.P.M. Numerical examples illustrating the fact that a quotient of two approximate numbers, both given as correct to k significant digits, may be correct to k or $(k-1)$ significant digits. For example, $a = 42.3 \pm 0.05$, and $b = 12.2 \pm 0.05$. Then (see section A)

$$\frac{a}{b} = 3.47 \pm 0.02, \quad \text{and} \quad \frac{a}{b} = 3.5.$$

Also, $a = 42.3 \pm 0.05$, and $b = 32.2 \pm 0.05$. Then

$$\frac{a}{b} = \frac{42.3}{32.2} \pm \frac{0.05(42.3 + 32.2)}{1{,}036.84} = 1.314 \pm 0.0036,$$

and $\frac{a}{b} = 1.31$.

Numerical examples illustrating the fact that a quotient of two approximate numbers, one given as correct to k significant digits, and the other given as correct to $(k+r)$, $r = 1, 2, 3, \ldots$, significant digits, may be correct to k or $(k-1)$ significant digits. For example, $a = 576.43 \pm 0.005$, and $b = 1.22 \pm 0.005$. Then

$$\frac{a}{b} = \frac{576.43}{1.22} \pm \frac{0.005 \cdot 577.65}{1.4884} = 472.\bar{5} \pm 1.9,$$

and $\frac{a}{b} = 470$.

Also, $a = 576.43 \pm 0.005$, and $b = 2.22 \pm 0.005$. Then

$$\frac{a}{b} = \frac{576.43}{2.22} \pm \frac{0.005 \cdot 578.65}{4.9284} = 259.7 \pm 0.6,$$

and $\frac{a}{b} = 260$.

D. THE RULES FOR DIVISION OF APPROXIMATE NUMBERS

I.P.M. The statement and illustration of the rules for division of two approximate numbers in terms of significant digits.

i) If two approximate numbers are given such that one is correct to k significant digits, and the other is correct to $(k+r)$, $r = 0, 1, 2, 3, \ldots$, significant digits, then if the product of the first significant digits on the extreme left of the approximate numbers is equal to or is greater than 10, or if the product of the first two significant digits on the extreme left of each of the approximate numbers is equal to or is greater than 1,000, the quotient of the two approximate numbers is correct to k significant digits. Otherwise the quotient is correct to $(k-1)$ significant digits.

ii) If one of the approximate numbers is correct to k significant digits, and the other approximate number is correct to $(k + r)$, $r = 2, 3, 4, \ldots$, significant digits, then the second approximate number should be rounded off to $(k + 1)$ significant digits.

T.R. The teacher should refer to the remarks on pages 57 and 58. For example find the ratio $\frac{a}{b}$, when $a = 17.36$, and $b = 31.45$.

$$\frac{a}{b} = 0.552 \text{ (the error is 0.0003).}$$

These remarks are very important, and they refer to division as well as multiplication.

E. THE DETERMINATION OF THE NUMBER OF SIGNIFICANT DIGITS OF TERMS OF QUOTIENTS REPRESENTED BY MATHEMATICAL CONSTANTS SUCH AS π, $\sqrt{2}$, $\sqrt{3}$, and $\sqrt{5}$

I.P.M. Numerical illustrations showing that certain numerical constants, such as π, $\sqrt{2}$, $\sqrt{3}$, and $\sqrt{5}$, may be determined to any number of significant digits, and that these constants should be used in accordance with the rules stated in section D above.

F. THE QUOTIENT OF AN EXACT NUMBER AND AN APPROXIMATE NUMBER

I.P.M. Illustrative examples showing that when one of the terms of a quotient is represented by an exact number and the other is approximate, the same rules hold; that is, the apparent and the relative errors of the exact number are zero, and the apparent error of the quotient is obtained in the same manner as above.

V. RAISING TO A POWER (pages 165–170)

The derivation of rules for raising to a power should follow the method of development that was employed in the case of multiplication. It is possible, however, to use a shorter procedure by basing all further discussions on the fact that in the case of raising to a power all factors are equal. All further remarks concerning raising to a power are similar in nature and content to the remarks made in reference to multiplication.

A. THE FORMULA FOR THE APPARENT ERROR OF A SQUARE OF AN APPROXIMATE NUMBER

I.P.M. The square of the approximate number $a = A + e_1$ is

$$a^2 = A^2 + 2Ae_1 + e_1^2.$$

Disregarding e_1^2, we have $a^2 = A^2 + 2Ae_1$, and the apparent error of the square of the approximate number is $2Ae_1$.

Numerical examples and problems leading to the calculation of the square of an approximate number and its apparent error. For example, $(42.5 \pm 0.05)^2 = 1{,}806.25 \pm 4.25$, and the square of the approximate number is 1,810.

T.R. If we use the expression for the apparent error of the product of two approximate numbers, that is, $Ae_2 + Be_1$, and if $A = B$, and $e_1 = e_2$, we get the apparent error of the square of an approximate number $2Ae_1$.

B. THE RELATIVE ERROR OF A SQUARE OF AN APPROXIMATE NUMBER

I.P.M. The relative error of the square of an approximate number is

$$\frac{2Ae_1}{A^2} = 2\frac{e_1}{A},$$

that is, it is twice the relative error of the approximate number itself.

Numerical examples and problems leading to the calculation of the relative error of the square of an approximate number.

C. THE SIGNIFICANT DIGITS OF A SQUARE OF AN APPROXIMATE NUMBER

I.P.M. Numerical examples illustrating the fact that a square of an approximate number given as correct to k significant digits may be correct to k or $(k - 1)$ significant digits. The determination of the number of significant digits of the square of an approximate number by means of the apparent error of the square.

D. THE RULE FOR SQUARING APPROXIMATE NUMBERS

I.P.M. The statement and illustration of the rule for squaring approximate numbers in terms of significant digits. For example, if an approximate number given as correct to k significant digits is less than $10^m \cdot 3.1623 \ldots$, that is, if the digits appear in the approximate number in this order, then its square is correct to $(k - 1)$ significant digits. Otherwise, its square is correct to k significant digits.

T.R. The development of this material should follow the procedure used in multiplication. The number $10^m \cdot 3.1623 \ldots$ represents the square root of 10^{2m+1} (m is an integer $\lesseqgtr 0$).

E. THE CUBE OF AN APPROXIMATE NUMBER

I.P.M. The cube of an approximate number $a = A + e_1$ is

$$a^3 = A^3 + 3A^2e_1 + 3Ae_1^2 + e_1^3.$$

Disregarding $3Ae_1^2 + e_1^3$, we get $a^3 = A^3 + 3A^2e_1$, and the apparent error is $3A^2e_1$, whereas the relative error is

$$3\frac{e_1}{A},$$

that is, the relative error of a cube of an approximate number is three times as large as the relative error of the approximate number itself. Numerical problems leading to the calculation of the cube of an approximate number and its errors.

T.R. For example, find the volume of a cube whose edge is $a =$ 8.3 in. $V = (8.3 \pm 0.05)\text{in}^3 = (572 \pm 11)\ \text{in}^3$, or $V = 600.\ \text{in}^3$ See section A above for an alternate procedure for the derivation of the formula for an apparent error of a cube of an approximate number.

F. THE TABLE OF SQUARES. INTERPOLATION. THE TABLE OF CUBES

I.P.M. Use of table of squares. Interpolation by means of proportional parts. For example, $(56.1)^2$ is obtained as follows:

$$\begin{array}{lr} & 57^2 = 3{,}249 \\ & 56^2 = 3{,}136 \\ \hline \text{difference} & 113; \end{array}$$

the difference for the increase of 0.1 is 11.3. Then $(56.1)^2$ is 3,147.3, and the result correctly stated is 3,150.

Use of tables of cubes and interpolation as above.

T.R. Tables of squares of numbers are practical if arranged in a two-way form as in Table 1. The use of tables of this form introduces

TABLE 1

	0	1	2	3	4	5	6	7	8	9
0	0	1	4	9	16	25	36	49	64	81
1	100	121	144	169	196	225	256	289	324	361
2	400	441	484	529	576	625	676	729	784	841
3	900	961	1024	1089	1156	1225	1296	1369	1444	1521
4	1600	1681	1764	1849	1936	2025	2116	2209	2304	2401
5	2500	2601	2704	2809	2916	3025	3136	3249	3364	3481
6	3600	3721	3844	3969	4096	4225	4356	4489	4624	4761
7	4900	5041	5184	5329	5476	5625	5776	5929	6084	6241
8	6400	6561	6724	6889	7056	7225	7396	7569	7744	7921
9	8100	8281	8464	8649	8836	9025	9216	9409	9604	9801
10	10000									

the pupils to a type of table that they will use frequently as they progress in their studies of mathematics. It is essential, however, that before the pupils begin to use tables arranged in a two-way form they should be given a knowledge of how to locate points in a co-ordinate plane when the two co-ordinates of the points are given. The teacher should show the pupils that the above table can be used to obtain squares of pure decimals and mixed decimals as well. He should show that if the decimal point is transposed in the argument (the number) a certain number of places, the result will be the displacement of the decimal point in the square twice the number of places in the same direction. For example, $22^2 = 484$; $2.2^2 = 4.84$; $0.22^2 = 0.0484$. The teacher should point out to the pupils that in the process of interpolation they cannot expect to secure any more significant digits in the square of an approximate number than the number of significant digits warranted by the approximate number itself. Moreover, if the table is arranged for numbers in the range 0–99, then the process of interpolation cannot be carried to more than one extra digit in the argument, that is, this table can give squares of approximate numbers for the range 0–999. In the case of the table of cubes of numbers which can be arranged in the same manner as the table of squares, the teacher should show the pupils that such a table can be used for cubes of pure decimals as well as mixed decimals. He should show that the transposition of the decimal point in the argument (the number) a certain number of places results in the displacement of the decimal point in the cube three times the number of places in the same direction.

VI. Extraction of Roots (pages 180–189)

The extraction of roots is the only numerical process in which numerical results of greater accuracy than the accuracy of the given data can at times be obtained.[*] This operation is the only one which is an exception to the principle of Approximate Computation that the accuracy of a numerical result should not exceed the approximativeness of the least accurate number with which certain numerical processes are performed. This property of root extraction becomes obvious when the relative error of a root of an approximate number is compared with the relative error of the approximate number itself.

---

[*] Except in cases such as $1{,}000 \pm \pi$, where 1,000 is an exact number, and π is numerically expressed by an approximate number.

A. THE FORMULA FOR THE APPARENT ERROR OF THE SQUARE ROOT OF AN APPROXIMATE NUMBER

I.P.M. Let the approximate number be $a = A + e_1$. Then

$$\sqrt{a} = \sqrt{A + e_1} = B + e_2,$$

where $B = \sqrt{A}$. Squaring the above equality we get:

$$A + e_1 = B^2 + 2Be_2 + e_2^2.$$

Then

$$A + e_1 = A + 2\sqrt{A}e_2 + e_2^2.$$

Disregarding e_2^2, we get: $e_1 = 2\sqrt{A}e_2$, or

$$e_2 = \frac{e_1}{2\sqrt{A}}.$$

For example, $\sqrt{42.5}$ is obtained as follows. The apparent error of 42.5 is ±0.05. The apparent error of the square root is

$$\frac{0.05}{2\sqrt{42.5}} = \frac{0.05}{2 \cdot 6.52} = 0.004 \text{ (approximately)}.$$

The square root, stated correctly, is 6.52.

T.R. Another method of presenting the apparent error of the square root of an approximate number is as follows: Let

$$\sqrt{A} + e_1 = \sqrt{A} + \frac{e_1}{2\sqrt{A}},$$

and to verify this fact we square both sides of the equality. We have then

$$A + e_1 = A + e_1 + \frac{e_1^2}{4A}.$$

Disregarding the term with e_1^2, we obtain an identity. Then

$$\frac{e_1}{2\sqrt{A}}$$

is the apparent error of $\sqrt{a}$.

B. THE RELATIVE ERROR OF THE SQUARE ROOT OF AN APPROXIMATE NUMBER

I.P.M. The relative error of the square root of an approximate number $a = A + e_1$ is

$$\frac{e_1}{2\sqrt{A}} : \sqrt{A} = \frac{1}{2}\frac{e_1}{A}.$$

In other words, the amount of error per unit of measure in terms of which the square root of an approximate number is expressed is half the amount of error per unit of measure in terms of which the approximate number itself is expressed. For example, $\sqrt{53.7}$ has a relative error equal to half the relative error of 53.7. We have: the relative error of 53.7 is $\frac{00.5}{53.7} = 0.00093$; the relative error of $\sqrt{53.7}$ is $\frac{1}{2}(0.00093) = 0.00047$.

Numerical examples and problems illustrating and making use of the above fact.

T.R. The fact stated in this section should be illustrated with numerical examples showing that in some cases the square root of an approximate number may have one more significant digit than the approximate number itself, for example, $\sqrt{934.}$ The apparent error of the square root is

$$\frac{0.5}{2\sqrt{934.}} = \frac{0.5}{2 \cdot 30.56} = \frac{0.5}{61.12} = 0.008.$$

The square root is then 30.56, correct to four significant digits.

The extraction of a square root should be preceded by the calculation of its apparent error, so that the extraction of the root is carried to one extra place, and then the resultant number is rounded off to the required number of significant digits as determined by its apparent error. The remarks of this section apply to section C below.

C. THE DETERMINATION OF THE SIGNIFICANT DIGITS IN THE SQUARE ROOT OF AN APPROXIMATE NUMBER

I.P.M. The use of the apparent error for the purpose of determining the number of significant digits in the square root of an approximate number.

The use of the relative error for the purpose of calculating the apparent error of the square root of an approximate number. For example, the apparent error of $\sqrt{934.}$ is calculated as follows. The relative error of 934. is

$$\frac{0.5}{934.} = 0.00054.$$

The relative error of $\sqrt{934.}$ is $\frac{1}{2}(0.00054) = 0.00027$. The first two significant digits of $\sqrt{934.}$ are 30, and hence the apparent error of $\sqrt{934.}$ is $0.00027 \cdot 30 = 0.008$. See examples in sections A and B above.

D. THE CUBE ROOT OF AN APPROXIMATE NUMBER

I.P.M. Let the approximate number be $a = A + e_1$. Then

$$\sqrt[3]{a} = \sqrt[3]{A + e_1} = B + e_2,$$

where $B = \sqrt[3]{A}$, or $B^3 = A$, and $B^2 = \sqrt[3]{A^2}$.

Then

$$A + e_1 = B^3 + 3B^2e_2 + 3Be_2^2 + e_2^3,$$

or

$$A + e_1 = A + 3\sqrt[3]{A^2}e_2 + 3\sqrt[3]{A}e_2^2 + e_2^3.$$

Disregarding the terms with e_2^2 and e_2^3, we get

$$e_1 = 3\sqrt[3]{A^2}e_2.$$

The apparent error e_2 of $\sqrt[3]{a}$ is then

$$e_2 = \frac{e_1}{3\sqrt[3]{A^2}}.$$

The relative error of $\sqrt[3]{a}$ is

$$\frac{e_1}{3\sqrt[3]{A^2}} : \sqrt[3]{A} = \frac{1}{3}\,\frac{e_1}{A}.$$

For example, $\sqrt[3]{85.4}$ is obtained as follows: The relative error of 85.4 is

$$\frac{0.05}{85.4} = 0.000586.$$

The relative error of $\sqrt[3]{85.4}$ is $\frac{1}{3}(0.000586) = 0.000195$. The first significant digit on the extreme left of $\sqrt[3]{85.4}$ is 4.; then the apparent error of $\sqrt[3]{85.4}$ is

$$4 \cdot 0.000\text{[illegible]}95 = 0.0008 \text{ (approximately)}.$$

Extracting the cube root (either by means of logarithms or by means of tables) we get

$$\sqrt[3]{85.4} = 4.4037,$$

or rounded off to thousandths, the cube root, correct to four significant digits, is 4.404.

T.R. See section A above for an alternate procedure in the derivation of the expression for the apparent error of the cube root of an approximate number $a = A + e_1$. Let

$$\sqrt[3]{A} + e_1 = \sqrt[3]{A} + \frac{e_1}{3\sqrt[3]{A^2}}.$$

Cubing both sides of the equality we get

$$A + e_1 = A + e_1 + \frac{e_1^2}{3A} + \frac{e_1^3}{27A^2}.$$

Dropping the terms with e_1^2 and e_1^3, we obtain an identity.

E. TABLES OF SQUARE ROOTS. INTERPOLATION. TABLES OF CUBE ROOTS

I.P.M. Interpolation by means of proportional parts. For example, $\sqrt{73.2}$ is obtained as follows: We determine the apparent error of the square root (see A and C above); it is 0.003. We then have:

$$\begin{array}{lr} & \sqrt{74} = 8.602 \\ & \sqrt{73} = 8.544 \\ \hline \text{difference} & 0.058 \end{array}$$

The difference for 0.2 is 0.012 which is to be added to $\sqrt{73}$. Then for $\sqrt{73.2}$ we have

$$8.544 + 0.012 = 8.556.$$

But the apparent error of the square root, which is 0.003, is in the thousandths. Hence, the square root, stated correctly to three significant digits, is 8.56.

Use of tables of cube roots.

T.R. Tables of square roots of numbers are best arranged in a two-way form as in Tables 2 and 3.

Table 2 gives roots of numbers having an odd number of significant digits to the left of the decimal point. Table 3 gives roots of numbers having an even number of significant digits to the left of the decimal point. The teacher should show the pupils that the transposition of the decimal point $2n$ places in the arguments of the table of square roots results in the displacement of the decimal point n places in the same direction in the square root itself.

The process of interpolation should be employed with great care. The table of square roots gives square roots to four significant digits (as a matter of fact we can construct a table of square roots that will give the roots to any number of digits, but its value will be questionable because the number of significant digits in a square root of an approximate number depends on the number of significant digits in

TABLE 2

	0	1	2	3	4	5	6	7	8	9
1.	1.000	1.049	1.095	1.140	1.183	$1.22\bar{5}$	$1.26\bar{5}$	1.304	1.342	1.378
2.	1.414	1.449	1.483	1.517	1.549	1.581	1.612	1.643	1.673	1.703
3.	1.732	1.761	1.789	1.817	1.844	1.871	1.897	1.924	1.949	$1.97\bar{5}$
4.	2.000	$2.02\bar{5}$	2.049	2.074	2.098	2.121	$2.14\bar{5}$	2.168	2.191	2.214
5.	2.236	2.258	2.280	2.302	2.324	2.345	2.366	2.387	2.408	2.429
6.	2.449	2.470	2.490	2.510	2.530	2.550	2.569	2.588	2.608	2.627
7.	2.646	$2.66\bar{5}$	2.683	2.702	2.720	2.739	2.757	$2.77\bar{5}$	2.793	2.811
8.	2.828	2.846	2.864	2.881	2.898	2.915	2.933	2.950	2.966	2.983
9.	3.000	3.017	3.033	3.050	3.066	3.082	3.098	3.114	3.130	3.146

TABLE 3

	0.	1.	2.	3.	4.	5.	6.	7.	8.	9.
1	3.162	3.317	3.464	3.606	3.742	3.873	4.000	4.123	4.243	4.359
2	4.472	4.583	4.690	4.796	4.899	5.000	5.099	5.196	5.292	5.385
3	5.477	5.568	5.657	$5.74\bar{5}$	5.831	5.916	6.000	6.083	6.164	$6.24\bar{5}$
4	$6.32\bar{5}$	6.403	6.481	6.557	6.633	6.708	6.782	6.856	6.928	7.000
5	7.071	7.141	7.211	7.280	7.348	7.416	7.483	7.550	7.616	7.681
6	7.746	7.810	7.874	7.937	8.000	8.062	8.124	8.185	8.246	8.307
7	8.367	8.426	8.485	8.544	8.602	8.660	8.718	$8.77\bar{5}$	8.832	8.888
8	8.944	9.000	9.055	9.110	9.165	9.220	9.274	9.327	9.381	9.434
9	9.487	9.539	9.592	9.644	9.695	9.747	9.798	9.849	9.899	9.950
10	10.000									

the approximate number itself). The process of interpolation will not result in any additional significant digits. Moreover, the process of interpolation should be employed as a substitute for the actual extraction of a root. The final number of significant digits in the square root of an approximate number is determined by means of the apparent error of the square root.

The table of cube roots is arranged in the same manner as the table of square roots, that is, in a two-way form. But in the case of a table of cube roots the transposition of the decimal point $3n$ places in the argument results in the displacement of the decimal point in the cube root n places in the same direction.

The determination of the number of significant digits in the cube root of an approximate number is effected by means of the apparent error of the cube root (see D above). Since it involves too much arithmetical work in the direct calculation of the apparent error of the cube root of an approximate number, the product of the relative error of this cube root and the first (or the first two) significant digit of the root (taken from the table) will give the first significant digit on the left of the apparent error of this root (Chap. V, sec. VI, B).

CHAPTER VII

APPROXIMATE COMPUTATION: LOGARITHMS OF NATURAL NUMBERS

THE proof that the numerical values of logarithms of natural numbers (excepting logarithms of numbers of the form b^m, where b is the base and m is an integer) are approximate is too difficult for the level of secondary school mathematics. The demonstration of the fact that numerical values of logarithms of numbers are generally approximate and that logarithms of approximate numbers are always approximate is, therefore, out of the question at the level of elementary algebra and numerical trigonometry. The teacher can, however, successfully and convincingly present this fact to the pupils so that they will learn it intuitively. It is needless to point out here that logarithms with the base 10 are used for the purpose of computation.

1. THE APPROXIMATE NATURE OF THE NUMERICAL VALUES OF LOGARITHMS OF NATURAL NUMBERS AND ESPECIALLY OF APPROXIMATE NUMBERS

I.P.M. The development of the notion of a logarithm as the exponent to which 10 is to be raised in order to obtain a given number. The approximativeness of the exponents of 10 for all numbers located in positions intermediate to 10^m (m being an integer).

T.R. In order to convince the pupils that the numerical values of logarithms of all numbers (except those of the form 10^m, where m is an integer) and especially of logarithms of all approximate numbers are approximate, the teacher should present the sequence of integral powers of 10, the exponents of these powers being the logarithms with the base 10. The intermediate values for the exponents, corresponding to the numbers located between the values of numbers 10^m (m being an integer), cannot be integral. At this stage of presentation the teacher should use a table of logarithms. The pupils will see for themselves that all one can ever expect from the numerical value of a logarithm of such an intermediate number is that it should be approximate.[1]

---

[1] For example the sequences of logarithms to the base 10: $\log 1 = 0$, $\log 10 = 1$,

II. THE RELATION OF THE APPROXIMATIVENESS OF LOGARITHMS AND THEIR CORRESPONDING APPROXIMATE NUMBERS

I.P.M. Inspection of tables of logarithms in order to determine the relationship between the number of significant digits of an approximate number and the number of digits in the mantissa of the logarithm corresponding to that number.

T.R. The relation between the approximativeness of the numerical value of a logarithm and the approximativeness of its corresponding number cannot be rigorously demonstrated at the level of secondary school mathematics. The teacher should resort to the same procedure and the same method of demonstration employed in the case of illustrating the fact that the numerical values of logarithms are generally approximate, that is, to the inspection of tables of logarithms of natural numbers. This inspection will reveal that in the case of numbers given as correct to three significant digits two places to the right of the decimal point of the mantissas of the logarithms remain fixed. As the third (the one on the extreme right) significant digit of the number changes, the third digit to the right of the decimal point of the mantissa changes also. It follows then that for approximate numbers given as correct to three significant digits the mantissas of their logarithms should have three digits, for approximate numbers given as correct to four significant digits the mantissas of their logarithms should have four digits and so on. (*Note:* No mention of whether the digits of the mantissas are significant or not should be made.)

III. TABLES OF LOGARITHMS. INTERPOLATION

I.P.M. Problems involving the use of logarithms for the purpose of computation. The calculation of areas of rectangles, circles, triangles; the calculation of volumes of rectangular solids, of a sphere, and cylinder.

T.R. Tables of logarithms should be arranged in a two-way form, and for numbers from 1,000 to 10,000 there should be four digits in the mantissas. For illustrations of the process of interpolation the reader is referred to pages 65 and 70.

---

log 100 = 2, log 1000 = 3. The mantissas of logarithms of all numbers located between 1 and 10, between 10 and 100, between 100 and 1000, and so forth, between 10^m and 10^{m+1} are represented by non-terminating decimal fractions. If feasible, an illustration of mantissas with 20 decimal places obtained from Logarithmica Britannica might be effective.

In order to distinguish between logarithms of exact numbers and logarithms of approximate numbers the following notation is suggested. A logarithm of an approximate number will be written with a suffix "a" as follows: "loga." Thus, when we write loga 64.75 we know that 64.75 is an approximate number, whereas log 3235 will denote that 3235 is an exact number.

The teacher should not fail to emphasize the fact that at no time can one obtain by means of interpolation from four-place mantissas, numbers with five significant digits or mantissas with five places.

All numerical processes with logarithms should be governed by the rules for Approximate Computation. In a great many cases a pupil starting out with logarithms whose mantissas have four places will obtain a logarithm of a final numerical result whose mantissa will have three places only. The teacher should stress the necessity of checking the approximativeness of numerical results by means of the apparent errors. This check will lead to the same result, and the pupils should see for themselves that in all cases, except the extraction of roots, the final numerical result of the performed numerical processes will be less accurate than the least accurate of the given numerical data. The process of obtaining a number by means of its logarithm is similar in procedure. The remarks concerning the relation between the number of significant digits of an approximate number and the number of places in the mantissa of its logarithm hold here also.

CHAPTER VIII

APPROXIMATE COMPUTATION: NATURAL TRIGONOMETRIC FUNCTIONS

THE introduction of trigonometric functions should emphasize the approximate nature of the numerical values of these functions. The proof of this nature is beyond the scope of difficulty of secondary school mathematics, but this consideration should not prevent the teacher from presenting the truth to the pupils. First of all, the teacher should avoid the definition of trigonometric functions as mere numbers with which one can perform certain short cuts in mensuration, so to speak. Trigonometric functions are much more than numbers. Besides expressing certain relationships, their numerical values, being ratios, represent measurements. It is not the purpose of this work to discuss the method of development of the topic of trigonometric functions at the level of elementary algebra and numerical trigonometry as such. The aim of this topic, within the scope of this work, is to develop the correct presentation of natural trigonometric functions (that is, their numerical values from the point of view of Approximate Computation) and to show as well as to determine the limitations that the nature of these functions imposes on their numerical values.

1. THE APPROXIMATE NATURE OF THE NUMERICAL VALUES OF NATURAL TRIGONOMETRIC FUNCTIONS

I.P.M. Numerical examples showing that the ratios representing natural trigonometric functions (i.e., their values) are ratios of numerical values of measures (i.e., of approximate numbers). Hence these ratios are approximate only.

T.R. The approximate nature of trigonometric functions should be presented from the point of view of measurement. On the level of elementary algebra and numerical trigonometry, that is, where the work is confined to right triangles only, natural trigonometric functions are introduced as ratios of two measures, and it is, therefore, consistent to stress this fact in order to introduce the notion of approximativeness of the numerical values of natural trigonometric functions. At the same time, the teacher should not fail to emphasize the fact that numerical values of measured angles, or angles computed by means of trigonometric functions, can be stated only approximately.

The few exceptions, that is, the values of trigonometric functions which are exact (if the angles are exact), should be pointed out to the pupils. These are:

sin 0°, sin 30°, and sin 90°,
cos 0°, cos 60°, and cos 90°,
tan 0°, tan 45°, and tan 90° (which is indeterminate).

II. THE RELATION BETWEEN THE APPROXIMATIVENESS OF AN ANGLE AND ITS CORRESPONDING NATURAL TRIGONOMETRIC FUNCTIONS

I.P.M. Examination of tables of natural trigonometric functions in order to determine the relation between the approximativeness of angles and their corresponding natural trigonometric functions.

T.R. The theoretical exposition of the connection between the approximativeness of an angle and the approximativeness of its corresponding trigonometric functions is too difficult for the level of secondary school mathematics. The pupils can, however, examine a table of natural trigonometric functions arranged at intervals of 6′ or 1′, and they can see for themselves that the number of significant digits in the numerical value of a trigonometric function is directly connected with the number of significant digits in the numerical value of an angle (expressed in total minutes). The only exception would be when the angles are either very small or very close to 90°, that is, when the functions are either irregular or insensible, or both (*irregular* means when the function varies too rapidly; *insensible*, when the variation of the value of the function is imperceptible for short intervals; see pages 244–245). The pupils will not fail to notice that as an angle is increased (or decreased) by 1′, two or three digits remain fixed in the numerical value of its corresponding trigonometric function, whereas the third or fourth digit changes according to the nature of the trigonometric function. It follows then that for approximate angles given as correct to 1′, natural trigonometric functions should be given as correct to three or four significant digits (four preferably). This is a new approach to numerical trigonometry and care should be exercised in presenting it properly, especially when practical field work is a part of the course.

III. THE USE OF NUMERICAL VALUES OF NATURAL TRIGONOMETRIC FUNCTIONS IN CONNECTION WITH OTHER APPROXIMATE NUMBERS

I.P.M. Solution of right triangles. Application of the rules for multiplication and division of approximate numbers.

T.R. The number of significant digits that should be used in natural trigonometric functions should be determined according to the rules for multiplication and division of approximate numbers. It should be understood, however, that for angles given as correct to 1′, four significant digits constitute the limiting approximativeness of a natural trigonometric function.

All work in numerical trigonometry should be governed by the rules for Approximate Computation stated in Chapter VI.

IV. TABLES OF NATURAL TRIGONOMETRIC FUNCTIONS. INTERPOLATION

I.P.M. Problems in calculating numerical values of natural trigonometric functions as well as calculating of angles corresponding to given natural trigonometric functions.

T.R. Tables of natural trigonometric functions for angles given as correct to 1° but allowing interpolation to 0.1° should be given with three significant digits. For angles of less than 10° the natural sine, cosine, and tangent may have two significant digits. The tables should be arranged in a two-way form.

Tables of natural trigonometric functions for angles given as correct to 1′ (or 6′, or 10′, but allowing at the same time to interpolate for angles up to the limit of precision, i.e., 1′) should be given with four significant digits.

For angles up to 16°39′ the natural sine, cosine, and tangent may have three significant digits. The tables should be arranged in a two-way form.

The process of interpolation was illustrated on page 65.

The teacher should not fail to emphasize the fact that at no time can these tables give results correct to fractions of a minute, nor should the pupils expect to secure by means of interpolation a numerical value of a natural trigonometric function correct to more than four significant digits.

In order to distinguish between trigonometric functions of exact angles and trigonometric functions of approximate angles the following notation is suggested. A natural trigonometric function of an approximate angle will be written with a suffix "a" as follows:

sina, cosa, tana, cota, seca, coseca.

If we write sina 32°10′, cosa *A*, tana *C*, we know that the angles 32°10′, *A*, *C* are approximate, whereas sin 25°20′, cos *B*, cot *D* will denote that the angles 25°20′, *B*, *D* are exact. Similarly in the case of logarithmic trigonometric functions of approximate angles we will write log sina, log cosa, log tana, log cota, log seca, log coseca. Thus log sina 42°15′, log tana *A*, log seca *B* will denote that the angles 42°15′, *A*, *B* are approximate, whereas log cos 30°, log tan *C*, log cot *D* will denote that the angles 30°, *C*, *D* are exact.

V. DETERMINATION OF AN APPROXIMATE ANGLE WHOSE NATURAL TRIGONOMETRIC FUNCTION IS GIVEN

I.P.M. Solution of right triangles for the purpose of finding angles. The use of the sine, cosine, and tangent formulas.

T.R. The work with natural trigonometric functions, whose numerical values are represented by approximate numbers, should be governed as stated above by the rules for Approximate Computation. It follows, then, that in a great number of cases natural trigonometric functions may be obtained, after computations are performed, with three or even two significant digits The number of significant digits in an approximate angle, which is to be obtained by means of its natural trigonometric function, should be determined as follows:

i) For angles up to and including 16°39′ the natural sine, cosine, and tangent may have three significant digits. If two significant digits are obtained in the natural trigonometric functions (sine and tangent), then the natural sine and tangent will give angles correct to 10′ only

ii) For all other angles, three significant digits in a natural sine, cosine, and tangent will give angles correct to 10′ only. In the case of two significant digits in the natural trigonometric functions, the angles will be correct to 1° only.

CHAPTER IX

APPROXIMATE COMPUTATION: APPROXIMATE COMPUTATION OF THE FIRST TYPE

COMPUTATION with approximate numbers should always lead to the question of the approximativeness of the final numerical result. The teacher should emphasize the fact that it is the aim of Approximate Computation to have the numerical processes performed in such a manner that the optinum possible information is secured in the final numerical result of performed computations. This is Approximate Computation of the first type.

I. THE NOTION OF OPTIMUM APPROXIMATIVENESS

I.P.M. Review exercises involving the processes illustrated in Chapters V, VI, and VII.

T.R. The notion of optimum approximativeness should be developed in the earliest stages of teaching Approximate Computation. The pupils should be shown that since they work with numbers which represent limited information, it is important to perform the required computations in such a manner that all the originally given information is utilized according to the rules governing Approximate Computation.

II. FORMULAS (pages 193–198)

The solution of problems in which approximate numbers are involved and in which formulas are used should not represent a merely mechanical process which leads to the performance of certain numerical computations. The pupils should be able to recognize the nature of the formulas used, their possibilities and their limitations. The nature of a formula is twofold. From the abstract point of view, a formula is a statement of relationship between certain things. From the numerical point of view, a formula is a statement of rules for computation that must be performed with numerical data in order to obtain a desired numerical result. A formula as a statement of relationship between objects may be of one of the three types that will be discussed below.

A. EXACT MATHEMATICAL FORMULAS

I.P.M. Formulas involving geometric relationships, such as

$$C = 2\pi r,\ A = \pi r^2,\ V = \frac{4}{3}\pi r^3,\ A = b^2,\ V = e^3,$$

$$A = \tfrac{1}{2}bh,\ A = bh,\ V = lwh,$$

perimeter = sum of all sides of a polygon,

$$a^2 + b^2 = c^2 \text{ (theorem of Pythagoras),}$$

$$\sin A = \frac{a}{c},\quad \cos A = \frac{b}{c},\quad \tan A = \frac{a}{b}$$

Time-distance formula: $s = vt$ in all of its variations.
Simple interest formula: $I = prt$.

T.R. Exact mathematical formulas are most commonly used in elementary algebra and numerical trigonometry. They represent facts of relationship derived from a purely mathematical procedure. These formulas are statements of geometric or algebraic relationship. But, even though they are mathematically exact, many of these formulas never give exact numerical results. If the numerical data to which they are to be applied are approximate, the numerical results are invariably approximate. This fact must be emphasized in order that no wrong impression may be created; exact mathematical formulas seldom lead to exact numerical results.

B. APPROXIMATE MATHEMATICAL FORMULAS

I.P.M. Illustration of the fact that interpolation of logarithmic and trigonometric functions, as well as interpolation of tables of squares, cubes, square roots, and cube roots, is not an exact process.

T.R. The use of approximate mathematical formulas is not very extensive in elementary mathematics. It cannot be denied, however, that such formulas find place and application as early as the study of elementary algebra and numerical trigonometry. Interpolation by proportional parts represents the application of an approximate mathematical formula. Even if the error of interpolation represented by the remainder term of the formula is not computed by the pupils in the process of interpolation, the teacher should not fail to show that such a remainder can be computed, and that in the case of a strict account of approximativeness this remainder plays an important part. It should be the aim of the teacher to emphasize the fact that interpolation is an approximate process unless it is carried far enough to the order of differences which vanish, and that logarithmic and trigonometric functions do not belong to the type that allows exact interpolation. Naturally, all this is far beyond the level of elementary algebra and numerical trigonometry; but mention of the fact that interpolation is an approximate process should be made in order that the students will not get the impression that it is an exact process.

C. SCIENTIFIC FORMULAS

I.P.M. The law of the pendulum: $t = 2\pi\sqrt{\frac{l}{g}}$; Boyle's law: $pv = k$; the law of acceleration: $a = \frac{1}{2}gt^2$; and the like.

T.R. A group of formulas, known as scientific formulas, the expression of which has been obtained either experimentally or from analysis of observed data, find a place in numerical work of elementary algebra. These formulas express physical relationships only approximately. They are not approximate mathematical formulas, however, because they represent a state of relationship different from that expressed by approximate mathematical formulas. Because of limitations of experimental apparatus, because of imperfections caused by the introduction of personal element in scientific observations, and because of the approximativeness of scientific methods, a true statement of relationship expressing the laws of Nature is still an unattainable goal. Numerical work with scientific formulas can never yield exact final numerical results, even though the data are exact. This fact should be emphasized by the teacher so that the pupils may carry it over into their school work in science.

D. THE APPARENT AND RELATIVE ERRORS OF A FORMULA

I.P.M. Evaluation of formulas and their apparent and relative errors.

T.R. Each formula applied to approximate numerical data leads to a certain apparent error of the final numerical result. This apparent error is obtained according to the rules for the respective numerical processes stated in Chapters V and VI. The evaluation of the apparent error which is of importance in elementary algebra and numerical trigonometry only as far as the determination of the last place of significance on the extreme right of the final numerical result, is therefore an important step in Approximate Computation.

In all numerical work in which the last operations are neither addition nor subtraction, it is not only more effective to use the relative errors, but by means of the relative errors much numerical work can be eliminated. The reader is referred to the analysis of multiplication and division in Chapter VI, from which he will observe that in the case of expressions involving all processes except addition and (or) subtraction, the relative error of the final numerical result may equal the sum of the relative errors of the individual terms of a formula. The apparent error of the final numerical result is obtained by means of the product of the first few significant digits on the left of the final numerical result and the relative error of this result. In order to facilitate the numerical work certain short cuts are very useful. These will be illustrated by the following example. Evaluate

$$\frac{23.52 \cdot 647.3}{3.142}$$

This can be rewritten in standard form, that is, so that each number will have only one significant digit to the left of the decimal point:

$$\frac{2.352 \cdot 10 \cdot 6.473 \cdot 10^2}{3.142}$$

The relative errors of the three numbers are:

of 23.52	0.00021
of 647.3	0.00008
of 3.142	0.00016
the relative error of the result:	0.00045

By means of short calculation we find that

$$\frac{2.3 \cdot 6.4}{3.1} = 4.6 \text{ (approximately).}$$

Then, for the apparent error of the result we have:

$$4.6 \cdot 10^3 \cdot 0.00045 = 2.17;$$

in other words, the final numerical result has an error in the units. We have then: $23.52 \cdot 647.3 = 15{,}224.496$, and

$$\frac{15{,}224.496}{3.142} = 4{,}845.$$

Hence, the final numerical result is $48\bar{5}0$.

the first two significant digits on the extreme left of each of the approximate numbers is $\geqslant 1{,}000$. Otherwise, the quotient of the two approximate numbers is correct to $(k - 1)$ significant digits.

7. The apparent error of the quotient of two approximate numbers is determined by means of the formula[5]

$$z > y_1 10^{s}|\wp_{(a:b)}|.$$

8. The resultant quotient of two approximate numbers should be rounded off in order that the apparent error of this quotient will be of the form

$$|\alpha| \leqslant 0.5 \cdot 10^{p-n-m+1},$$

m being the number of significant digits in the resultant quotient.

III. Extraction of Roots

Let A be the exact value of the approximate number a. The absolute error of a is α, and the relative error of a is $\wp_a$. Then

$$|a| \leqslant |A| + |\alpha|,$$
$$|\alpha| = |A - a| < \epsilon_1, \quad (\epsilon_1 \to 0).$$

We shall consider the square root first. We have then

$$\sqrt{A - \alpha} < \sqrt{a} < \sqrt{A + \alpha}.$$

Also,

$$\sqrt{a} \leqslant \sqrt{A \pm \alpha}.$$

The absolute error of the square root of an approximate number a is

$$|\sqrt{A} - \sqrt{a}| \geqslant |\sqrt{A} - \sqrt{A \pm \alpha}|,$$

or

$$|\sqrt{A} - \sqrt{A \pm \alpha}| = \frac{|A - (A \pm \alpha)|}{|\sqrt{A} + \sqrt{A \pm \alpha}|}.$$

Then

$$\frac{|\alpha|}{|\sqrt{A} + \sqrt{A \pm \alpha}|} \leqslant \frac{|\alpha|}{2|\sqrt{A}|}.$$

Since

$$\alpha + \epsilon_2 = \alpha_1, \quad (\epsilon_2 \to 0),$$
$$\lim_{\epsilon_2 \to 0} \frac{|\alpha + \epsilon_2|}{|2\sqrt{A}|} \leqslant \frac{|\alpha_1|}{2|\sqrt{A}|}.$$

Since

$$0 < A_2 < A < A_1$$
$$\frac{|\alpha_1|}{2|\sqrt{A}|} < \frac{|\alpha_1|}{2|\sqrt{A_2}|}.$$

---

[5] See Chap. XIV, sec. xv.

The expression for the apparent error of the square root of an approximate number is identical in form and nature with the expression for the absolute error of the square root of an approximate number. The direction of the absolute (apparent) error of the square root of an approximate number is the same as the direction of the absolute (apparent) error of the approximate number itself.

The above method may be extended to the case of the nth ($n = 3, 4, 5, \ldots$) root of an approximate number a. We have

$$\sqrt[n]{A - \alpha} < \sqrt[n]{a} < \sqrt[n]{A + \alpha},$$

and

$$\sqrt[n]{a} \lesseqgtr \sqrt[n]{A \pm \alpha}.$$

The absolute error of the nth root of the approximate number a is

$$\left|\sqrt[n]{A} - \sqrt[n]{a}\right| \geqslant \left|\sqrt[n]{A} - \sqrt[n]{A + \text{[illegible]}}\right|$$

from which we obtain the expression for the absolute error of the nth root

$$\frac{|\alpha|}{n\left|\sqrt[n]{A^{n-1}}\right|},$$

and, since

$$\alpha + \epsilon_2 = \alpha_1, \quad (\epsilon_2 \longrightarrow 0).$$

$$\operatorname*{Limit}_{\epsilon \to 0} \frac{|\alpha + \epsilon_2|}{\left|n\sqrt[n]{A^{n-1}}\right|} \leqslant \frac{|\alpha_1|}{n\left|\sqrt[n]{A^{n-1}}\right|}.$$

Also, since

$$0 < A_2 < A < A_1,$$

$$\frac{|\alpha_1|}{n\left|\sqrt[n]{A^{n-1}}\right|} < \frac{|\alpha_1|}{n\left|\sqrt[n]{A_2^{n-1}}\right|}.$$

The expression for the apparent error of the nth root of an approximate number a is identical in form and nature with the expression for the absolute error of the nth root of the approximate number.

The direction of the absolute (apparent) error of the nth root of an approximate number is the same as the direction of the absolute (apparent) error of the approximate number itself.

A. THE RELATIVE ERROR

The relative error of the square root of an approximate a, whose relative error is $\wp_a$, is

$$|\wp_{\sqrt{a}}| = \frac{|\alpha|}{2|\sqrt{A}|} \cdot \frac{1}{|\sqrt{A}|} = \frac{|\alpha|}{2|A|},$$

that is,

$$|\wp_{\sqrt{a}}| = \tfrac{1}{2}|\wp_a|.$$

It does not follow from the above, however, that

$$\wp'_{\sqrt{a}} = \tfrac{1}{2}\wp'_a;$$

the true relationship between the upper limit of the relative error of the square root of an approximate number and the upper limit of the relative error of the approximate number itself is

$$\wp'_{\sqrt{a}} \leqslant \tfrac{1}{2}\wp'_a.$$

The reader will find no difficulty in deriving the generalized expression for the relative error of the *n*th root of an approximate number *a*, which is

$$|\wp_{\sqrt[n]{a}}| = \frac{|\alpha|}{n|A|},$$

or

$$|\wp_{\sqrt[n]{a}}| = \frac{1}{n}\,|\wp_a|.$$

Also, by virtue of the results shown in Chapter XVI, sections A and C *iv*, we have

$$\wp'_{\sqrt[n]{a}} \leqslant \frac{1}{n}\,\wp'_a, \quad (n \leqslant 9),$$

and

$$\wp'_{\sqrt[n]{a}} < \frac{1}{n}\,\wp'_a, \quad (n > 9).$$

B. THE ABSOLUTE AND RELATIVE ERRORS

The use of the absolute (apparent) error alone in the case of the evaluation of the approximativeness of a root of an approximate number is impracticable from the point of view of the time and labor necessary. This can be readily observed from the formula for the absolute (apparent) error of the root.

The analysis of the approximativeness of a root of an approximate number consists of two problems. The first, and the essential part of the analysis, is the determination of the relative error of the root; and the second is the determination of the apparent error of the root by means of this relative error, and the first significant digit (or first few significant digits) of the root itself.

We shall consider in detail the case of the square root of an approximate number. The approximativeness of higher roots is determined by the general method indicated above.

In the process of the analysis of the approximativeness of a square root of an approximate number it is necessary to distinguish the cases when the approximate number is integral, a pure decimal, or a mixed decimal. We shall consider the case of an integral number in detail. The remaining two cases follow directly from the first one. Moreover, in the case of an integral number we shall have to distinguish two cases:

(1) When the number of the digits is $q = 2m + 1$.

(2) When the number of the digits is $q = 2m$.

Also, if $q > k$, k being the number of significant digits in the approximate number, then the approximate number may have one or several zeros on its extreme right, and this case will have to be considered.

Reference will be made to Table 7 which gives the relationship between q, the first significant digit on the extreme left of the approximate number, and the first significant digit of the square root of the approximate number.

TABLE 7

The value of q	The first significant digit on the extreme left of the approximate number a	The first significant digits on the extreme left of the square root
$2m + 1$	1, 2, and 3	1
	4, 5, 6, 7, and 8	2
	9	3
$2m$	10 to 15 inclusive	3
	16 to 24 inclusive	4
	25 to 35 inclusive	5
	36 to 48 inclusive	6
	49 to 63 inclusive	7
	64 to 80 inclusive	8
	81 to 99 inclusive	9

In the analysis developed in the following sections we shall consider the approximate number a of the form

$$a = 10^q(10^{k-1}x_1 + 10^{k-2}x_2 + \cdots + 10x_{k-1} + x_k + z10^{-1}),$$

where

$$q \text{ is an integer } \gtreqqless 0,$$

$$k \text{ is an integer } > 0, \quad |q| \gtreqqless k,$$

$$1 \leqslant |z| \leqslant 5,$$

$$x_1 \neq 0; \quad x_i = 0, 1, 2, \cdots, 9; \quad i = 2, 3, 4, \cdots, k.$$

In the case when A is integral, $q \geqslant 0$.

If $|q| = 2m$, then the square root of the approximate number a is

$$\sqrt{a} \approx 10^m \left\{ 10^{\frac{1}{2}(k-1)} y_1 + \cdots + \frac{z10^{-1}}{2(10^{\frac{1}{2}(k-1)} y_1 + \cdots)} \right\}.$$

If $|q| = 2m + 1$, then the square root of the approximate number a is

$$\sqrt{a} \approx 10^{m-1} \left\{ 10^{\frac{1}{2}k} w_1 + \cdots + \frac{z10^{-1}}{2(10^{\frac{1}{2}k} w_1 + \cdots)} \right\}.$$

In the actual numerical process, the above represents the breaking up of the extraction of the square root of an approximate number a into two parts.[6]

(1) The determination of the integral part of the square root of the approximate number.

(2) The determination of the "decimal" part, that is, the fractional part of the square root of the approximate number. In the process of root extraction, in this case, it is required that, should the approximate number a itself have a fractional part, this fractional part should always have an even number of digits. Should there be an odd number of digits in the fractional part of the approximate number a, that is, the approximate number a has an odd number of decimals, then an odd number of decimals (zeros) is added on the extreme right of this approximate number a, so that the number of decimals is made even.

If the integral part of an approximate number has $2s$ digits, then the integral part of the square root of this approximate number has s digits. If the integral part of the approximate number has $(2s + 1)$ digits, then the integral part of the square root of the approximate number has $(s + 1)$ digits.

Should there be no integral part in the true value of the approximate number a, that is, the true value of the approximate number is a "pure decimal," then by multiplying and dividing the approximate number by 10^m (m an even positive integer), the square root is extracted from the literal part of the approximate number.[7]

---

[6] See Chrystal, G.: Algebra; A. & C. Black, Ltd., London, 1926; Part I, Chap. XI, pp. 210–215.

i. The relative error

Let the approximate number a be correct to k significant digits. We shall denote the first significant digits on the left of the apparent error of the approximate number by z, z being equal to 1, 2, 3, 4, or 5. Then

$$|\wp_a| < \frac{z}{x_1 10^k},$$

and

$$|\wp_{\sqrt{a}}| < \frac{z}{2x_1 10^k}.$$

But then

$$\frac{\wp'_{\sqrt{a}}}{\wp'_a} \cong 10^{-r}, \quad (r = 0, 1),$$

that is, the relative error of the square root of an approximate number is either of the order "zero" or of the order "one" in comparison with the relative error of the approximate number itself. The reader will find no difficulty in considering the specific values of x_1 and the respective values of z and verifying the above result.

ii. The square root of an integral number

In the following analysis we shall consider $q = 0$; that is, when $k = 2m$,

$$\sqrt{a} = 10^{m-1}y_1 + 10^{m-2}y_2 + \cdots + \frac{z10^{-1}}{2(10^{m-1}y_1 + 10^{m-2}y_2 + \cdots)},$$

where $y_1 = 3, 4, 5, 6, 7, 8,$ or 9.

If $k = 2m + 1$,

$$\sqrt{a} = 10^m y_1 + 10^{m-1}y_2 + \cdots + \frac{z10^{-1}}{2(10^m y_1 + 10^{m-1}y_2 + \cdots)}$$

where $y_1 = 1, 2,$ or 3.

(1) *Let $k = 2m$.*

Then

$$\frac{z10^{-1}}{2(10^{m-1}y_1 + 10^{m-2}y_2 + \cdots)}$$

---

[7] *Ibid.*

occupies the $2m$th place in the expansion of $\sqrt{a}$. But

$$\frac{z}{2y_1}$$

by virtue of the values that z and y_1 may respectively take on, may be, and is <1. Hence, it occupies the $(2m+1)$st place in the expansion of $\sqrt{a}$. If $z = 1$, and $A \geqslant 25 \cdot 10^r$,

$$\frac{z}{2y_1} \leqslant 10^{-1},$$

and it occupies then the $(2m+2)$nd place in the expansion of $\sqrt{a}$, that is, there are then $(k+1)$ significant digits in $\sqrt{a}$.

(2) *Let* $k = 2m + 1$.

Then

$$\frac{z10^{-1}}{2(10^m y_1 + 10^{m-1}y_2 + \cdot\cdot\cdot)}$$

occupies the $(2m+2)$nd place in the expansion of $\sqrt{a}$. But

$$\frac{z}{2y_1}$$

by virtue of the values that z and y_1 may respectively take on, may be, and at times is, <1. This takes place, when

$$\left(\frac{z}{2}\right)^2 \leqslant A10^r.$$

The above results may be presented in tabular form. The square root of an approximate number, given as correct to $k = 2m + 1$ significant digits, will have the apparent error in the $(2m+3)$rd place of significance, that is, the square root will be correct to $(k+1)$ significant digits when

(1) $z = 1$;
(2) $z = 2$;
(3) $z = 3$; and $A \geqslant 225 \cdot 10^r$;
(4) $z = 4$; and $A \geqslant 4 \cdot 10^r$;
(5) $z = 5$; and $A \geqslant 625 \cdot 10^r$.

The above results may be verified by means of the relative error test. This is left to the reader.

Should $q > 0$, and $q = 2t$, then the same results follow as when $k = 2m$. The same also apply when $q = 2t + 1$ and $k = 2m + 1$. If $q = 2t$ and $k = 2m + 1$, or $q = 2t + 1$ and $k = 2m$, then the same results follow as when $k = 2m + 1$.

iii. The square root of a pure decimal and of a mixed decimal

If we apply the results obtained in section III, B, we may derive the criteria for the approximativeness of square roots of approximate numbers, when a is either a pure decimal or a mixed decimal.

When a is a pure decimal, $q < 0$, and $k \leqslant |q|$. If $k = 2m$ and $|q| = 2t$, or $k = 2m + 1$, and $|q| = 2t + 1$, then the case when $k = 2m$ and the number is integral $(q = 0)$ holds good. If $k = 2m$, and $|q| = 2t + 1$, or $k = 2m + 1$, and $|q| = 2t$, then the case applies as when the number is integral $(q = 0)$ and $k = 2m + 1$.

When a is a mixed decimal, $q < 0$, and $|q| < k$. If $k = 2m$ and $q = 2t$, or $k = 2m + 1$, and $|q| = 2t + 1$, then the case when $k = 2m$, and the number is integral $(q = 0)$ holds good. If $k = 2m$, and $|q| = 2t + 1$, or $k = 2m + 1$, and $|q| = 2t$, then the case when $k = 2m + 1$, and the number is integral $(q = 0)$ holds good.

C. SIGNIFICANT DIGITS

The results of the above sections may be translated in terms of significant digits. We shall retain the same notation as in the above section.

The number of significant digits in the approximate number is k $(k = 2m$, or $k = 2m + 1)$.

The number of the digits to the right of the decimal point of the approximate numbers in (b) and (c) below is q, $(|q| = 2t$, or $|q| = 2t + 1)$.

(a) In an integral number $q \geqslant 0$;
(b) in a pure decimal $q < 0$, and $|q| \geqslant k$;
(c) in a mixed decimal $q < 0$, and $|q| < k$.

1. If we have an approximate number a such that
   (a) it is an integral number;
   (b) it is a pure decimal;
   (c) it is a mixed decimal;

with $k = 2m$ significant digits and $|q| = 2t$, or $k = 2m + 1$ significant digits and $|q| = 2t + 1$, then we have the results in Table 8.

2. If we have an approximate number a such that
   (a) it is an integral number;
   (b) it is a pure decimal;
   (c) it is a mixed decimal;

with $k = 2m$ significant digits and $|q| = 2t + 1$, or $k = 2m + 1$ significant digits and $|q| = 2t$, then we have the results in Table 9.

TABLE 8

The significant digit z of the apparent error	The value of the approximate number a is	The number of significant digits in the square root
1	$A < 25 \cdot 10^r$	k
	$A \geqslant 25 \cdot 10^r$	$k + 1$
$2 \leqslant z \leqslant 5$	All values	k

TABLE 9

The significant digit z of the apparent error	The value of the approximate number a is	The number of significant digits in the square root
1	All values	$k + 1$
2	All values	$k + 1$
3	$A < 225 \cdot 10^r$	k
	$A \geqslant 225 \cdot 10^r$	$k + 1$
4	$A < 4 \cdot 10^r$	k
	$A \geqslant 4 \cdot 10^r$	$k + 1$
5	$A < 625 \cdot 10^r$	k
	$A \geqslant 625 \cdot 10^r$	$k + 1$

D. THE RULE FOR EXTRACTION OF ROOTS

Extraction of roots of an approximate number should be performed in accordance with the following rules:

1. The relative error of the nth root of an approximate number may be n times as small as the relative error of the approximate number itself.

2. The direction of the relative (apparent) error of the nth root of an approximate number is the same as the direction of the relative (apparent) error of the approximate number itself.

3. The number of significant digits in the square root of an approximate number a given as correct to k significant digits is given in Tables 8 and 9.

4. When an approximate number a is given as correct to k significant digits, and no information is given concerning the magnitude of the apparent (or relative) error of the approximate number, then

it should be assumed that the apparent error of the approximate number is ±0.5 of a unit to which the approximate number is correct, and the number of significant digits in the square root of this approximate number is determined by means of the tables given in section III, C.

5. The apparent error of the *n*th root of an approximate number is obtained by means of the formula[8]

$$z > y_1 10^m |\wp_{\sqrt[n]{a}}|.$$

6. The resultant *n*th root of an approximate number should be rounded off in order that the apparent error of the *n*th root may be of the form

$$|\alpha| \leqslant 0.5 \cdot 10^{\mu+n-m+1},$$

m being the number of significant digits to which the square root is correct.

---

[8] See Chap. XIV, sec. XV.

CHAPTER XVIII

APPROXIMATE COMPUTATION OF THE FIRST TYPE

(Concluded)

I. Approximate Computation with Exact Numbers

Very often the process of computation involves besides approximate numbers exact numbers also. The exact number can be thought of as an approximate number whose apparent (absolute) error and whose relative error are "zero." Therefore, the rules and formulas for computation with approximate numbers are applicable to the combinations of exact numbers and approximate numbers.

A. ADDITION AND SUBTRACTION

When adding or subtracting an exact number to or from an approximate number, or vice versa, the absolute (apparent) error of the sum or of the difference is equal to the sum of the absolute (apparent) errors of the two numbers (the absolute values of the errors being considered). It follows then that the absolute (apparent) error of the sum of the two numbers or of their differences is identical with the absolute value of the absolute (apparent) error of the approximate number.

If an exact number has significant digits to the right of the place of significance of the last significant digit on the extreme right of the approximate number, such digits have no significance in the resultant sum or difference. The proper procedure, therefore, is to round off the exact number. The exact number thus ceases to be exact, and becomes approximate. For example, consider the sum:

$$(23{,}450 + 3) + 124{,}763.$$

We have then:

$$\begin{array}{r} 23{,}450 + 3 \\ 124{,}760 + 3 \\ \hline 148{,}210 + 6 \\ \hline 148{,}220 - 4 \end{array}$$

Should the last significant digit on the extreme right of the exact number be to the left of the last place of significance on the extreme right of the approximate number, the approximate number should be rounded off to conform with the last place of significance of the exact number. For example:

$$(23{,}452. + 0.4) + 124{,}000.$$

We have then:

$$\begin{array}{r} 23{,}452. + 0.4 \\ 124{,}000. \\ \hline 147{,}452. + 0.4 \end{array}$$

The same procedure should be followed in subtraction.

B. MULTIPLICATION

The product of an exact number and an approximate number is invariably approximate.

Suppose we have an approximate number $|a| \leqslant |A| + |\alpha|$ and an exact number N. The product of the two numbers is

$$|N||a| \leqslant |N||A| + |N||\alpha|.$$

The relative error of the product is still $\wp_a$; that is, the product of an exact number and an approximate number is just as accurate as the approximate number which enters into this product as a factor. It may happen, however, that if the approximate number is correct to k significant digits, the product of an exact number and an approximate number may be correct to $(k-1)$, k, or $(k+1)$ significant digits. In the case of the product of an exact number and an approximate number all depends on the magnitude of the first significant digit on the extreme left of the resultant product and of the first significant digit on the left of the absolute (apparent) error of the resultant product. For example:

Approximate number	Apparent error	Relative error
43.2	+0.03	+0.00069
32.	+0.00	+0.00000
1,382.4	+0.96	+0.00069

The product, 1,383. + 0.4, is correct to four significant digits.

Approximate number	Apparent error	Relative error
43.2	−0.03	−0.00069
74.	+0.00	+0.00000
3,196.8	−2.22	−0.00069

The product, 3,190 + 4, is correct to three significant digits.

Approximate number	Apparent error	Relative error
43.2	+0.05	+0.00116
2.	+0.00	+0.00000
86.4	+0.1	0.00116

The product, 86. + 0.5, is correct to two significant digits.

Should the approximate number be correct to k significant digits, and the exact number contain $(k + r)$ digits $(r = 2, 3, 4, \ldots)$, the exact number should be rounded off to $(k + 1)$ significant digits. The exact number thus ceases to be exact and becomes approximate.

If the approximate number is given as correct to $(k + r)$ significant digits, and the exact number contains k digits, the approximate number should never be rounded off.

C. DIVISION

In the process of division of an approximate and an exact number two cases must be distinguished:

(1) The approximate number a is the dividend, and the exact number N is the divisor.

(2) The exact number N is the dividend, and the approximate number a is the divisor.

(1) If the quotient is $\frac{a}{N}$, then the absolute (apparent) error of the quotient is

$$\frac{|\alpha|}{|N|},$$

and the relative error of the quotient is $\wp_a$. For example:

Approximate number	Apparent error	Relative error
44.2	+0.04	+0.0009
12.	+0.00	+0.0000
3.683	+0.003	+0.0009

The quotient, 3.69 − 0.004, is correct to three significant digits.

Approximate number	Apparent error	Relative error
44.2	−0.01	−0.00023
12.	0.00	0.00000
3.6833	−0.0008	−0.00023

The quotient, 3.683 − 0.0005, is correct to four significant digits.

(2) If the quotient is $\frac{N}{a}$, then the absolute (apparent) error of the quotient is

$$\frac{|N||\alpha|}{|A^2|},$$

and the relative error of the quotient is $\wp_a$. For example: If the exact number is 342.. and the approximate number is 16.4 with an apparent error of -0.05 and a relative error of -0.00305, then we have 20.87 with an apparent error of -0.06 and a relative error of -0.00305. The quotient, $20.8 + 0.01$, is correct to three significant digits. Should the approximate number be correct to k significant digits, and the exact number contain $(k + r)$ digits $(r = 2, 3, \ldots)$, the exact number should be rounded off to $(k + 1)$ significant digits. The exact number thus ceases to be exact, and becomes approximate.

If the approximate number is given as correct to $(k + r)$ significant digits, and the exact number contains k digits, the approximate number should never be rounded off.

II. COMPLEX APPROXIMATE COMPUTATION

The formulas which are applied to numerical data (approximate or exact, or approximate and exact combined) call for a series of numerical operations. For example, the formula for the sides of a right triangle,

$$c = \sqrt{a^2 + b^2},$$

requires the following operations:

1. Raising to a power—twice.
2. Addition.
3. Extraction of a square root.

The "tangent formula" for a triangle whose three sides are given is

$$\tan\frac{A}{2} = \sqrt{\frac{(s-b)(s-c)}{s(s-a)}},$$

where $s = \frac{1}{2}(a + b + c)$, and a, b, and c represent the numerical values of the lengths of the sides of the triangle. The above formula involves the following operations:

1. Addition $(2s = a + b + c)$.
2. Division $(s = \frac{1}{2}(a + b + c))$.
3. Subtraction—three times.

4. Multiplication—twice.
5. Division.
6. Extraction of a square root.
7. The use of trigonometric tables.
8. Multiplication, in order to obtain the angle A.

Much of this work is eliminated by the use of logarithmic tables. We have then:

$$\log \tan \frac{A}{2} = \tfrac{1}{2}\left\{\log(s-b) + \log(s-c) - \log s - \log(s-a)\right\}.$$

This formula involves the following operations:

1. Addition ($2s = a + b + c$).
2. Division ($s = \frac{1}{2}(a + b + c)$).
3. Subtraction—three times.
4. The obtaining of logarithms.
5. Addition—twice.
6. Subtraction.
7. Division of the last difference (6) by 2.
8. The use of trigonometric tables.
9. Multiplication, in order to obtain the angle A.

Assuming that the entire numerical data to which a definite formula (mathematically exact, mathematically approximate, or scientific) is applied are approximate, the numerical result that is obtained after all the computations are performed is necessarily approximate. The approximativeness of the final numerical result is determined by the absolute error and the relative error tests.

The evaluation of the apparent (absolute) error and of the relative error of the final numerical result of the application of a formula to a given set of approximate numerical data is, therefore, an essential part of the process of Approximate Computation.

The analysis of the various numerical operations as developed in Chapters XVI and XVII shows that in certain processes one type of error is more essential and more effective, and in other processes the other type is more effective and essential. For example:

(1) In addition and subtraction the absolute (apparent) error is the best criterion for the study of the approximativeness of a numerical result of the performed operations.

(2) In multiplication, division, raising to a power, and extraction

of roots, the relative error is more effective than the absolute (apparent) error.

It does not follow, however, that one type of error is to be taken as entirely predominant to the exclusion of the other type. In addition and subtraction the question of the precision of a measure is more important than the question of the accuracy of the same measure, because only numerical data representing measures of the same precision should be added or subtracted. In the remaining four numerical processes—multiplication, division, raising to a power, and extraction of roots—the precision of the measure is of no essential importance. It is the accuracy of the measure that plays the important part in the process of the computation, and the numerical result is judged by the accuracy of the expression; whereas the precision of the numerical result, though information concerning this criterion is important, is secondary to the information concerning the accuracy of the same result.

We can formulate rules for the general procedure:

1. If the final numerical process of the application of a formula to the stated approximate numerical data is addition or subtraction, then the approximativeness of the final numerical result is judged by the absolute (apparent) error of this result, and the relative error is used as a check for the determination of the correctness and the approximativeness of the final numerical result. The relative error is subsidiary in this case.
2. If the final numerical process of the application of a formula to the stated approximate numerical data is either multiplication, division, raising to a power, or extraction of roots, then the approximativeness of the final numerical result is judged by the relative error of this result, and the absolute (apparent) error is used as a subsidiary and a check for the determination of the correctness and the approximativeness of the final result of the computations.
3. The final numerical result is to be stated with the absolute (apparent) error, relative error, or both, according to the requirements stated in advance.
4. If the formula that is applied to the stated numerical data (approximate or exact) is mathematically approximate, and the limits of the error of the formula are stated in the formula, this error should be taken into account and (usually) the apparent (absolute) error of the final numerical result must be adjusted to this error by adding to it the value of the error of the formula or subtracting this error

from it, as the case may be. The interpolation formula can be mentioned as a formula of this type.[1]

A. ABSOLUTE AND RELATIVE ERRORS OF AN EXACT MATHEMATICAL FORMULA

An exact mathematical formula, if applied to a set of stated approximate numerical data, invariably leads to an approximate numerical result. If an exact mathematical formula is applied to exact numerical data, and in the final form, after all the numerical substitutions have been made, it contains the following, then the numerical results of the computations that are performed in the process of the evaluation of the formula are invariably approximate.

1. Ratios that are represented by irreducible fractions whose denominators contain as factors powers of integers other than 2 and 5, and very often these are not excepted.

2. Irrational numbers such as π and e, logarithmic and trigonometric functions (with a few exceptions; see Chapters XX and XXI).

3. Powers of certain expressions, such as the expression in the compound interest formula, etc.

4. Surds, unless the number under the radical sign is a perfect power of an integral index, an integral multiple of the index of the surd.

It is therefore essential that whenever an exact mathematical formula, an approximate mathematical formula, or a scientific formula is applied to numerical data, whether approximate or exact, the absolute (apparent) error and the relative error of the final numerical result be evaluated. If the numerical data are exact, an examination of the formula and of the numerical data can immediately reveal the fact whether the final numerical result is approximate or exact. If the final numerical result is approximate, the computer should at once proceed to derive the expression for the absolute error and the relative error of the final numerical result, that is, the absolute error and the relative error of the formula. This will eliminate unnecessary and repetitious work in the process of the computation.

The expressions for the absolute error and the relative error of the formulas are obtained according to the rules stated in section II, and according to the results obtained in Chapters XVI and XVII.

(1) If the formula is of the form

$$v = ka \pm mb,$$

where k and m are certain constants, approximate or exact, whose exact values (in practice the true values) are denoted by the corresponding capital letters, and whose absolute errors (in practice the apparent errors) are denoted by α's with the respective subscripts; a and b are approximate. We have then:

---

[1] See Chap. XIV, sec. III, B.

$$|\alpha_v| = |A||\alpha_k| + |K||\alpha| + |B||\alpha_m| + |M||\beta|.$$

If

$$|\wp_{(ka)}| < |\wp_{(mb)}|,$$

then

$$|\wp_k| + |\wp_a| \leqslant |\wp_{(ka \pm mb)}| \leqslant |\wp_m| + |\wp_b|.$$

If

$$|\wp_{(ka)}| = |\wp_{(mb)}|,$$

then

$$|\wp_a| + |\wp_k| = |\wp_{(ka \pm mb)}| = |\wp_m| + |\wp_b|.$$

If

$$|\wp_{(ka)}| > |\wp_{(mb)}|,$$

then

$$|\wp_k| + |\wp_a| \geqslant |\wp_{(ka \pm mb)}| \geqslant |\wp_m| + |\wp_b|.$$

(2) If the formula is of the form

$$v = \frac{ab}{cd},$$

where a, b, c, and d are approximate, then

$$|\alpha_v| = \frac{|\alpha||BCD| + |\beta||ACD| + |\gamma||ABD| + |\delta||ABC|}{|CD|^2}.$$

and

$$|\wp_v| \leqslant |\wp_a| + |\wp_b| + |\wp_c| + |\wp_d|.$$

a, b, c, and d themselves may represent results of numerical processes of addition, subtraction, multiplication, division, raising to a power, extraction of a root, the taking of logarithms, trigonometric operations, or their combinations.

(3) If the formula is of the form

$$v = \sqrt[n]{\frac{ab}{cd}},$$

where a, b, c, and d are approximate, then

$$|\alpha_v| = \frac{|\alpha||BCD| + |\beta||ACD| + |\gamma||ABD| + |\delta||ABC|}{n(|CD|)^2 \sqrt[n]{\left(\frac{AB}{CD}\right)^{n-1}}},$$

and

$$|\wp_v| \leqslant \frac{1}{n}\left\{|\wp_a| + |\wp_b| + |\wp_c| + |\wp_d|\right\}.$$

It is obvious that in order to eliminate much labor, the computer can in such cases as (2) and (3) resort to the use of the relative error, and the absolute (apparent) error of the final numerical result can be obtained by means of the formula

$$z \geqslant y_1 10^k |\wp_r|,$$

where y_1 is the first significant digit on the extreme left of the final numerical result, $\wp_r$ is the relative error of this result, and z is the first significant digit on the left of the apparent error of the final numerical result. If several significant digits on the extreme left of the final numerical result are known, they should be used in the above formula. This will result in a more closely stated apparent error.

The direction of the apparent (absolute) error and of the relative error of the final numerical result of the application of a formula to a set of numerical data is determined according to the rules stated in Chapters XVI and XVII.

B. THE ORDER OF THE NUMERICAL OPERATIONS

The order of the numerical operations in the case of each particular formula depends on the type of the operation. It is important that the errors of the same type be grouped together as much as possible. In other words, addition and subtraction should be performed first and all the other operations should be performed afterwards, or vice versa. At all times the "mixing" of the various operations whose errors are of different types should be avoided as much as possible.

The least advantageous numerical process is that of subtraction. This is because of the property of the relative error of the difference of two approximate numbers a and b, $(|\wp_a| < |\wp_b|)$,

$$|\wp_a| < |\wp_{(a-b)}| \gtreqless |\wp_b|.$$

For example:

Approximate number	Apparent error	Relative error
984.362	−0.0003	−0.0000003
−977.683	+0.0004	+0.0000004
[illegible]679	−0.0007	−0.00015

The absolute value of the relative error of the difference is about 500 times as large as the absolute value of the smallest relative error of one of the approximate numbers. Rounding off the resultant difference and its errors we have

$$6.678 \qquad +0.0003 \qquad +0.000045$$

Therefore, whenever possible, the subtraction which would result in a very large relative error (in absolute value) should be avoided. This can be accomplished by changing subtraction into multiplication. This is done as follows:

Suppose that we have two approximate numbers

$$|a| \leqslant |A| + |\alpha|,$$

and

$$|b| \leqslant |B| + |\beta|,$$

such that the difference $(A - B)$ is exceedingly small in comparison with either A or B. Suppose that $B = X^3$ (or generally, X^n), and $A = (X + \Delta X)^3$. Then

$$A = X^3 + 3X^2\Delta X + 3X(\Delta X)^2 + (\Delta X)^3,$$

and

$$B = X^3.$$

Then

$$A - B = 3X^2\Delta X + 3X(\Delta X)^2 + (\Delta X)^3.$$

Now, if ΔX is very small, then $(\Delta X)^2$ and $(\Delta X)^3$ are much smaller in comparison with either A or B, and therefore we may disregard all the terms containing powers of ΔX. We have then

$$A - B \approx 3X^2\Delta X.$$

For example, consider the difference $(a - b)$, where

$$a \approx 6.2583 + 0.00002,$$
$$b \approx 6.2534 - 0.00003.$$

Ordinarily we would write $(a - b) \approx 0.0049 + 0.00005$. If we use the procedure outlined above, we have:

$$6.2583 + 0.00002 \approx (1.84283 + 0.000003)^3$$
$$6.2534 - 0.00003 \approx (1.84235 - 0.000002)^3$$

Then

$$\Delta X \approx 1.84283 - 1.84235 = 0.00048,$$

and

$$\begin{aligned} 3X^2\Delta X &\approx 3 \cdot 0.00048(1.84235 + 0.000002)^2, \\ &\approx 0.00144(3.39425 + 0.000007), \\ &\approx 0.00488772 + 0.00000001, \end{aligned}$$

and

$$(a - b) \approx 0.0048877 + 0.00000003.$$

The use of cubes is not absolutely necessary. The computer should determine the type of function that is most suitable for his purposes, as well as for the purpose of transforming subtraction into multiplication.

III. THE APPROXIMATIVENESS OF ROOTS OF EQUATIONS

Solutions of many problems are effected by means of equations. Naturally, the question concerning the approximativeness of the roots of equations must arise in the process of the solution of these equations.

If a given equation, or a given set of equations, is based on approximate data, the roots of such equations are necessarily approximate. Equations constructed on exact data in most of the cases (unless the equations are artificial) lead to solutions, that is, to roots which are approximate also.

In algebra, the following types of equations are usually solved:

1. Linear equations.
2. Simultaneous linear equations.
3. Quadratic equations.
4. Simultaneous quadratic equations.
5. Cubic equations and equations of higher degrees which are solved by special methods, such as Horner's, the method of iterations.
6. Simple exponential equations which are solved by means of logarithms.
7. Simple trigonometric equations which occur very often in trigonometry.

A. THE LINEAR EQUATIONS

These equations are of the form

$$ax + by + c = 0,$$

where a, b, and c are given constants (exact or approximate).

The solution of this type leads to a ratio, and in many cases, when a, b, c, and x (or y) are exact, such a ratio is represented by a non-terminating decimal. The numerical value of the ratio in such a case, that is, if it is represented by a non-terminating decimal, should never contain more significant digits than any of the given numbers a, b, c, and x (or y) with the least number of significant digits, unless a specific accuracy is prescribed in advance. In other words, the solution of the linear equation of the form

$$ax + by + c = 0$$

is generally at the most as accurate (or as precise) as the least accurate element entering into such an equation. If a, b, c, and x (or y) or at least one of them is approximate, the solution of the equation is invariably approximate. Naturally, the absolute (apparent) error and the relative error of the solution of this equation must be, therefore, calculated along the methods illustrated in the preceding sections.

B. SIMULTANEOUS LINEAR EQUATIONS WITH SEVERAL UNKNOWNS

These equations are of the form

$$\text{1.}\quad \begin{aligned} a_1x + b_1y &= c_1 \\ a_2x + b_2y &= c_2 \end{aligned}$$

$$\text{2.}\quad \begin{aligned} a_1x + b_1y + c_1z &= d_1 \\ a_2x + b_2y + c_2z &= d_2 \\ a_3x + b_3y + c_3z &= d_3 \end{aligned}$$

or generally

$$\text{3.}\quad \begin{aligned} a_1x + b_1y + \cdot\ \cdot\ \cdot + m_1w &= r_1 \\ a_2x + b_2y + \cdot\ \cdot\ \cdot + m_2w &= r_2 \\ a_3x + b_3y + \cdot\ \cdot\ \cdot + m_3w &= r_3 \\ \cdots\cdots\cdots\cdots\cdots\cdots & \cdots\cdots \\ \cdots\cdots\cdots\cdots\cdots\cdots & \cdots\cdots \\ a_mx + b_my + \cdot\ \cdot\ \cdot + m_mw &= r_m \end{aligned}$$

The equations of this type are usually solved by the method of the determinants. Assuming the equations are consistent, we have[2]

(1)
$$x = \frac{\begin{vmatrix} c_1 & b_1 \\ c_2 & b_2 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}}, \qquad y = \frac{\begin{vmatrix} a_1 & c_1 \\ a_2 & c_2 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}}.$$

(2)
$$x = \frac{\begin{vmatrix} d_1 & b_1 & c_1 \\ d_2 & b_2 & c_2 \\ d_3 & b_3 & c_3 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix}}; \qquad y = \frac{\begin{vmatrix} a_1 & d_1 & c_1 \\ a_2 & d_2 & c_2 \\ a_3 & d_3 & c_3 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix}};$$

$$z = \frac{\begin{vmatrix} a_1 & b_1 & d_1 \\ a_2 & b_2 & d_2 \\ a_3 & b_3 & d_3 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix}}.$$

---

[2] See Chrystal, G.: Algebra, Part I; A. & C. Black, Ltd., London, 1926; pp. 352–395.

(3)
$$x = \frac{\begin{vmatrix} r_1 & b_1 & c_1 & \cdots & m_1 \\ r_2 & b_2 & c_2 & \cdots & m_2 \\ r_3 & b_3 & c_3 & \cdots & m_3 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ r_m & b_m & c_m & \cdots & m_m \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 & c_1 & \cdots & m_1 \\ a_2 & b_2 & c_2 & \cdots & m_2 \\ a_3 & b_3 & c_3 & \cdots & m_3 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_m & b_m & c_m & \cdots & m_m \end{vmatrix}}; \text{ etc.}$$

The solutions of the sets of simultaneous linear equations are usually obtained in the form of ratios, and very often these ratios are represented by non-terminating decimals. Even if the coefficients a_i, b_j, c_k, ... are exact in their values, and the above ratios are represented by non-terminating decimals, and others by reducible fractions which can be represented by terminating decimals, the roots of such equations are approximate in their values. If one of the coefficients is approximate, then the roots of the equations are invariably approximate. It is, therefore, necessary to evaluate the absolute (apparent) error and the relative error of these roots. If the coefficients a_i, b_j, c_k, ... are exact, the evaluation of these errors is impossible until the accuracy is determined. If the coefficients are approximate these errors can be obtained according to the methods illustrated in the preceding sections. It can be stated that the roots of the equations are at the most as accurate as the least accurate coefficient of one of the unknowns.

C. THE QUADRATIC EQUATION

Let the quadratic equation be of the form

$$ax^2 + bx + c = 0,$$

where the coefficients a, b, and c are either exact or approximate. The roots of this equation are

$$r_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

Even if the coefficients a, b, and c are exact, the roots of the quadratic equation may be approximate. First of all $(b^2 - 4ac)$ may not be a perfect square (we assume that the roots of the equation may be real or imaginary). Secondly, even if $(b^2 - 4ac)$ is a perfect square, the final ratio may be represented by a non-terminating decimal only.

If the coefficients a, b, and c are exact, and the roots of the equation are approximate, the evaluation of the errors, both the absolute (apparent) error and the relative error, becomes possible only after the accuracy to which the roots are to be calculated is determined. It should be understood, however, that the calculation of such roots should be carried to the extent of the accuracy that is essentially required by the limitations of the problem in question. There is no sense in calculating the roots of the quadratic equation to, say five places of accuracy, when the exact data are given to one or two digits. No general rule can be formulated in this case. The computer should use common sense in carrying out the calculations.

If the coefficients a, b, and c of the quadratic equation

$$ax^2 + bx + c = 0$$

are approximate, then the roots of the quadratic equation are invariably approximate in their numerical values.

The approximativeness of the roots of the quadratic equation, all (or at least one) of whose coefficients of the unknowns are approximate, can be predetermined in advance. This, however, does not eliminate the necessity of the evaluation of the absolute (apparent) errors and the relative errors of the roots of the equation. From the well-known property of the roots of the equation we have that

$$r_1 + r_2 = -\frac{b}{a},$$

$$r_1 r_2 = \frac{c}{a}.$$

Since we have the sum of the roots and the product of the roots, and since the roots are approximate, it follows that the roots should, in the case of the sum, have absolute errors (apparent errors) of the same order, that is, the roots should be of the same order of precision; and in the case of the product of the roots, these roots should have relative errors at least of the same order, or at the most one of the relative errors should be of the order "one" in comparison with the other relative error. Since we have two quotients,

$$-\frac{b}{a} \quad \text{and} \quad \frac{c}{a},$$

each of the quotients is at the most as accurate as the least accurate number a, b, or c. Therefore, the sum of the roots and the product of the roots of the quadratic equation are at the most as accurate as the least accurate coefficient of the equation.

Since, in the process of addition, if the absolute (apparent) errors of the approximate numbers are equal to their upper limits, or even greater than their upper limits (this may happen to one absolute error if the directions of the absolute errors are opposite), there may be a loss of one significant digit on the extreme right of the resultant sum of the two approximate numbers. In the process of multiplication there may be a loss of one significant digit on the extreme right of the resultant product of two approximate numbers, and a similar loss may take place in the case of division. It follows, therefore, that the roots of the quadratic equation are as accurate as the least accurate (sometimes as precise as the least precise, depending on the nature of the problem) coefficient *a*, *b*, or *c* of the quadratic equation.

D. SIMULTANEOUS QUADRATIC EQUATIONS

The solution of simultaneous quadratic equations is based on the methods illustrated in the preceding paragraphs. It follows, therefore, that in most of the cases, even if the coefficients of the unknowns are exact, the roots of the equations are approximate. If the coefficients of the unknowns are all (or at least one of them is) approximate, then the roots of the equations are invariably approximate.

The extent of the approximativeness of the roots of the simultaneous quadratic equations cannot be generally predetermined for all the cases, for there are different forms of simultaneous quadratic equations. It can be stated, however, that the roots of the simultaneous quadratic equations, whose coefficients are all (or at least one of them) approximate, are as accurate and as precise (or as precise) as the least accurate (or as the least precise) of the coefficients of the unknowns of the equations.

E. THE CUBIC EQUATION

Every cubic equation can be reduced to the form[3]

$$x^3 + ax + b = 0.$$

If we denote

$$P = \frac{b}{2} + \sqrt{\frac{b^2}{4} + \frac{a^3}{27}}; \quad Q = \frac{b}{2} - \sqrt{\frac{b^2}{4} + \frac{a^3}{27}},$$

then one of the roots of the cubic equation is

$$-\sqrt[3]{P} - \sqrt[3]{Q}.$$

---

[3] See Cajori, Florian: An Introduction to the Modern Theory of Equations; Macmillan Co., New York, 1924; pp. 68–73. Chrystal, G.: Algebra, Part I; A. & C. Black, Ltd., London, 1926; pp. 549, 550.

This last expression may give some information concerning the approximativeness of the root of the cubic equation, if the coefficients a and b are both (or at least one of them is) approximate.

Both P and Q are at the most as precise as $\frac{b}{2}$ or

$$\sqrt{\frac{b^2}{4}+\frac{a^3}{27}}$$

Caution should be exercised in the case of Q on account of the subtraction. The above square root is, at the most, as accurate as (and probably less accurate than) either $\frac{b}{2}$ or $\frac{a}{3}$. Hence, the root of the cubic equation is at the most as accurate as the least accurate coefficient of the unknowns, a or b.

The approximativeness of the roots of the cubic equation can be analyzed from another angle. Again, from the Theory of Equations we have the following property of the roots of a cubic equation

$$ax^3+bx^2+cx+d=0,$$

where a, b, c, and d are approximate, or at least one of them is. Let the roots of the cubic equation be r_1, r_2, and r_3. Then

$$r_1+r_2+r_3=-\frac{b}{a},$$

$$r_1r_2+r_1r_3+r_2r_3=\frac{c}{a},$$

$$r_1r_2r_3=-\frac{d}{a}.$$

The sum of the roots is at the most as precise as the quotient $\frac{b}{a}$. It is, therefore, as accurate as this quotient. The sum of the products of the roots taken two at a time is as precise and as accurate as the quotient $\frac{c}{a}$. The product of the three roots is as accurate as the quotient $\frac{d}{a}$. The sum of three approximate numbers may lose one significant digit on the extreme right. The product of three approximate numbers is at the most as accurate as the least accurate factor, but it may have a relative error which is of the order "one" in comparison with the relative error of the least accurate factor. The sum of three products of three approximate numbers taken two at a time is as precise as the least precise product. The quotient of two approximate numbers is at the most as accurate as the least accurate dividend or divisor. It follows, therefore, that the roots of the equation are at the most as accurate as the least accurate coefficient. Since there are processes of division and multiplication in this analysis, the accuracy criterion should predominate. This remark applies to the analysis of the quadratic equation. Generally, in the solution of equations the roots should be computed to significant digits.

The above analysis is only a generalization. At all times the absolute (apparent) error and the relative error of the roots should be obtained, and these criteria should be taken as the basis of the evaluation of the approximativeness of the roots of the equation.

Although all the coefficients of the cubic equation should be exact, the roots of the equation may very often be approximate. It may be said that in exceptional cases (if the equations are artificial) the roots of a cubic equation are exact. If the roots of the cubic equation, whose coefficients are exact, are approximate, the absolute (apparent) errors and the relative errors of the respective calculated roots can be determined only after the degree of the approximativeness of the computed roots is decided upon by the computer.

F. EXPONENTIAL EQUATIONS

The exponential equations in the simplest form are

$$a^x = b.$$

The solution of this equation is

$$x = \frac{\log b}{\log a}.$$

Both log a and log b are generally approximate (unless a and b are integral powers of 10). It is understood that the logarithms are taken to the base 10.

The solution of the exponential equation is, therefore, a ratio, and, generally, the numerator and the denominator of this expression are both approximate. It follows then that the solution of the exponential equation is generally approximate, whether a and b are both (or whether one of them is) approximate or exact.

The extent of the approximativeness of the solution of an exponential equation of the above type and form depends on the extent of the approximativeness (accuracy) of the logarithms, and as it will be shown in the analysis of the logarithmic functions, the precision of the logarithms is closely related to the accuracy of the approximate number to which the logarithm corresponds. We conclude, therefore, that the solution of an exponential equation is at the most as accurate as the approximate numbers which enter as elements into the equation,

$$a^x = b;$$

that is, the numerical value of x can be expected to be at the most as accurate as the least accurate approximate number a or b.

Should numbers a and b both be exact, the accuracy of the solution of the exponential equation can be determined from the point of view of the accuracy of the logarithms of these two numbers, and the numerical value of x should at no time contain more significant digits than the number (a or b) or its logarithm with the least number of significant digits.

G. TRIGONOMETRIC EQUATIONS

The extent to which trigonometric equations find a place in the courses of elementary trigonometry is very limited, but the question concerning the approximativeness of the solutions of such equations should be considered. We shall illustrate one of such equations.

Let the equation be of the form

$$a\cos x + b\sin x = c.$$

Divide both sides of the equation by $\sqrt{a^2 + b^2}$. We have then

$$\frac{a}{\sqrt{a^2+b^2}}\cos x + \frac{b}{\sqrt{a^2+b^2}}\sin x = \frac{c}{\sqrt{a^2+b^2}}.$$

Let

$$\tan k = \frac{b}{a},$$

then

$$\frac{a}{\sqrt{a^2+b^2}} = \cos k; \quad \frac{b}{\sqrt{a^2+b^2}} = \sin k.$$

The above equation is then rewritten as

$$\cos k \cdot \cos x + \sin k \cdot \sin x = \frac{c}{\sqrt{a^2+b^2}},$$

or

$$\cos(x - k) = \frac{c}{\sqrt{a^2+b^2}}.$$

If

$$|c| < \left|\sqrt{a^2 + b^2}\right|,$$

then

$$x - k = \text{arc cos } \frac{c}{\sqrt{a^2 + b^2}},$$

and

$$x = \text{arc cos } \frac{c}{\sqrt{a^2 + b^2}} + \text{arc cos } \frac{a}{\sqrt{a^2 + b^2}}.$$

The accuracy (or precision) of the solution of this trigonometric equation depends on the accuracies (and precisions) of the three numbers a, b, and c. We conclude then that the solution of this equation is invariably approximate, because all the trigonometric functions (with a few exceptions only; see Chapter XXI) are represented by numerical values which are approximate.

The absolute (apparent) error of the solution of the above equation may be equal in absolute value to the sum of the absolute values of the absolute (apparent) errors of the two arc cosines.

The analysis of the approximativeness of the trigonometric functions and the determination of the relation between the approximativeness of the numerical value of an angle and the approximativeness of the numerical values of its corresponding trigonometric functions are found in Chapter XXI.

IV. APPROXIMATE COMPUTATION WITH APPROXIMATE NUMBERS WHICH ARE GIVEN AS CORRECT WITHIN A RELATIVE ERROR

In the analysis of the various numerical processes that was made in Chapters XVI and XVII constant reference was made to the relative error, and its relation to the absolute error, as well as to the function of the relative error in the process of Approximate Computation and the final evaluation of the approximativeness of the final numerical result of the Approximate Computation. It was shown that in all the numerical processes, except in the processes of addition and subtraction, the relative error tests are the more effective and the more direct means of the evaluation of the approximativeness of the numerical results; the function of the absolute error, however, was not at the same time discarded. The absolute error played a subsidiary part whenever the relative error was the effective criterion of the approximativeness of the numerical data.

We shall assume that the approximate numerical data to which formulas (mathematically exact, mathematically approximate, or scientific) are to be applied are given within a stated relative error. In order not to repeat the development of the preceding chapters, we shall briefly state the general procedure when the relative error only is given.

A. ADDITION

The relation between the relative error of the sum of several approximate numbers and the relative errors of these respective numbers is

$$|\wp_a| < |\wp_{\Sigma a}| < |\wp_n|,$$

where $\wp_a$ is the smallest (in absolute value) relative error, and $\wp_n$ is the greatest (in absolute value) relative error.

The relative error of the sum of several approximate numbers cannot be numerically determined from the set of the given relative errors of the respective approximate numbers. It is, therefore, necessary to obtain the relative error of the sum of the approximate numbers by means of the formula

$$|\wp_{\Sigma a}| = \frac{|\alpha| + |\beta| + |\gamma| + \cdots + |\nu|}{|A| + |B| + |C| + \cdots + |N|}.$$

The respective absolute (apparent) error of the approximate numbers can be obtained by means of the formula

$$z > x_1 10^k |\wp_a|.$$

B. SUBTRACTION

The relation between the relative error of the difference of two approximate numbers and the relative errors of these two approximate numbers is

$$|\wp_a| < |\wp_{(a-b)}| \gtreqless |\wp_b|,$$

and obviously, since this relation is indefinite from the purely numerical point of view, the magnitude of the relative error of the difference of the two approximate numbers a and b cannot be determined from it.

In order that we may be able to compute the relative error of the difference of two approximate numbers, the absolute (apparent) errors of these two approximate numbers must be obtained, and this is effected by means of the formula

$$z > x_1 10^k |\wp_a|,$$

and the relative error of the difference of the two approximate numbers is obtained by means of the formula

$$|\wp_{(a-b)}| = \frac{|\alpha| + |\beta|}{|A - B|}.$$

C. MULTIPLICATION

The relative error of the product of several approximate numbers a, b, c, ... , etc., whose relative errors are given as $\wp_a$, $\wp_b$, $\wp_c$, . . . , etc., is obtained by means of the formula

$$|\wp_{(abc\ldots n)}| \leqslant \sum_{k=a}^{n} |\wp_k|,$$

and the absolute (apparent) error of the product of these approximate numbers (which should by all means be calculated) is obtained by means of the formula

$$z > y_1 10^k |\wp_{(abc\ldots n)}|.$$

The use of the absolute (apparent) error in the process of multiplication is not necessary; this error is of importance in the final stage of the process of multiplication as an aid to the relative error for the purpose of the evaluation of the approximativeness of the resultant product from the point of view of the "significant digits approximativeness."

D. DIVISION

The relative error of the quotient of two approximate numbers a and b, whether a is the dividend or the divisor, is obtained by means of the formula

$$|\wp_{(a:b)}| \leqslant |\wp_a| + |\wp_b|,$$

where $\wp_a$ and $\wp_b$ are the relative errors of the respective approximate numbers.

The absolute (apparent) error of the resultant quotient is obtained by means of the formula

$$z > y_1 10^k |\wp_{(a:b)}|.$$

The use of the absolute (apparent) error in the process of division is not necessary; this error is of importance in the final stage of the process of division as an aid to the relative error for the purpose of the evaluation of the approximativeness of the resultant product from the point of view of the "significant digits approximativeness."

E. THE DIRECTION OF THE ERRORS

The numerical processes of addition, subtraction, multiplication, and division should be carried out with approximate numbers whose errors are of the same direction. Should the approximate numbers be given with relative errors whose directions are opposite (except in subtraction and division), it will be necessary to transform the approximate numbers so that the directions will become the same.

This transformation is accomplished by means of the absolute (apparent) errors. It is, therefore, necessary to obtain the apparent errors of the approximate numbers, and then to change the limits of the true values of the approximate numbers whose apparent errors are of opposite directions. In the case of the process of addition it is immaterial whether we use the upper limits or the lower limits of the true values (although the lower limits are of less advantage). In the case of multiplication, however, the upper limits of the true values of the approximate numbers are preferable. After the limits of the true values of the approximate numbers have been changed, and the new apparent (absolute) errors have been obtained, their corresponding relative errors should be calculated.

F. RAISING TO A POWER

The relative error of the nth power of the approximate number a ($n = 2, 3, 4, \ldots$), whose relative error is $\wp_a$, is obtained by means of the formula

$$|\wp_{(a^n)}| = n|\wp_a|, \quad (n = 2, 3, 4, \cdots),$$

and the absolute (apparent) error of the nth power of the approximate number is obtained by means of the formula

$$z > y_1 10^k |\wp_{(a^n)}|.$$

The use of the absolute (apparent) error in the process of raising to a power is not necessary; this error is of importance in the final stage of the process of raising to a power as an aid to the relative error for the purpose of the evaluation of the approximativeness of the resultant nth power of the approximate number from the point of view of the "significant digits approximativeness."

G. EXTRACTION OF ROOTS

The relative error of the nth root of the approximate number a ($n = 2, 3, 4, \ldots$), whose relative error is $\wp_a$, is obtained by means of the formula

$$|\wp_{\sqrt[n]{a}}| = \frac{1}{n}\,|\wp_a|, \quad (n = 2, 3, 4, \cdots),$$

and the absolute (apparent) error of the *n*th root of the approximate number is obtained by means of the formula

$$z > y_1 10^k |\wp_{\sqrt[n]{a}}|.$$

The use of the absolute (apparent) error in the process of extraction of roots is not necessary; this error is of importance in the final stage of the process of extraction of roots as an aid to the relative error for the purpose of the evaluation of the approximativeness of the resultant *n*th root of the approximate number from the point of view of the "significant digits approximativeness."

V. APPROXIMATE COMPUTATION WITH APPROXIMATE NUMBERS WHICH ARE GIVEN AS CORRECT WITHIN A STATED NUMBER OF SIGNIFICANT DIGITS

If a set of stated numerical data to which a formula (mathematically exact, mathematically approximate, or scientific) is to be applied is given as approximate and as correct to a stated number of significant digits only, and it is impossible to determine the extent of the magnitudes of the apparent error and of the relative error, it is the usual practice to assume that the apparent (absolute) error of any of the approximate numbers in such a case is equal in its absolute value to the upper limit of the absolute error, that is, to

$$10^{p+n-k+1} \cdot 0.5,$$

and since the directions of the errors are not known, it is assumed that the error is either positive or negative, and it is denoted, therefore, by ±.

The use of the significant digits as the only criterion of the approximativeness of the numerical data, thus avoiding the use of the absolute (apparent) error and the use of the relative error, should not be made a part of the general practice of Approximate Computation. In practical measurements the apparent errors of the measures can be determined at least to the first significant digit. If tables with mathematical constants are used, the apparent error should be determined also; and if necessary it should be given in the table.

VI. SUMMARY

The results that were obtained in the preceding chapters, as well as in this chapter, can be summarized as follows:

1. In addition and in subtraction the absolute (apparent) error is the basic criterion for the determination of the approximativeness of the resultant sum or of the resultant difference of the approximate numbers. Approximate numbers that are to be added or subtracted must be such that their absolute errors are of the same order.

2. In multiplication and in division the relative error is the basic criterion for the determination of the approximativeness of the resultant product or quotient, and the absolute (apparent) error is only a subsidiary aid to the relative error for the purpose of the evaluation of the approximativeness of the numerical result from the point of view of the "significant digits approximativeness." Both the relative error and the absolute (apparent) error are necessary for their combined use in the evaluation of the approximativeness of the numerical result of a computation. The apparent error of the final numerical result can be obtained by means of the relative error and of the first significant digit (or the first few) on the extreme left of the number representing the final numerical result. In the processes of multiplication and division the relative errors of the approximate numbers should be of the same orders, or at the most the relative errors of some of the approximate numbers may be of the order "one," but not higher in comparison with the relative error smallest in absolute value of some of the approximate numbers.

3. In raising to a power and in extraction of roots the relative error is the basic criterion for the determination of the approximativeness of the final numerical result, and the absolute (apparent) error is only a subsidiary aid to the relative error for the purpose of the determination of this approximativeness from the point of view of the "significant digits approximativeness." Both the relative error and the absolute (apparent) error are necessary, because their combined use in the evaluation of the approximativeness of the numerical result enables one to carry out this process effectively. The apparent error of the final numerical result can be obtained by means of the relative error and the first significant digit (or the first few) on the extreme left of the final numerical result.

4. The determination of the approximativeness of the final numerical result of the performed computations as to the number of

significant digits it may contain is effected by means of Table 10 which shows the relationship between the number of significant digits of an approximate number, the upper limit of the relative error of the approximate number, and the first significant digits on the ex-

TABLE 10

The upper limit of the relative error is located between	The number of the significant digits of the approximate number	
	The significant digit of the apparent error is $\leqslant x_1$	The significant digit of the apparent error is $> x_1$
0.5 and 0.1	–	1
0.1 and 0.01	1	2
0.01 and 0.001	2	3
0.001 and 0.0001	3	4
0.0001 and 0.00001	4	5
....	.	.
....	.	.
10^{-n} and $10^{-(n+1)}$	n	$n + 1$

TABLE 11

Operation	Absolute error	Relative error
$\sum^{n} a$	$\sum^{n} \vert\alpha\vert$	$\dfrac{\sum^{n} \vert\alpha\vert}{\sum^{n} \vert A\vert}$
$a - b$	$\vert\alpha\vert + \vert\beta\vert$	$\dfrac{\vert\alpha\vert + \vert\beta\vert}{\vert A - B\vert}$
$abc \cdots n$	$\sum \vert\alpha\vert \vert B\vert \vert C\vert \cdots \vert N\vert$ summed to n terms	$\sum_{k=a}^{n} \vert\wp_k\vert$
$\dfrac{a}{b}$	$\dfrac{\vert A\vert \vert\beta\vert + \vert B\vert \vert\alpha\vert}{\vert B\vert^2}$	$\vert\wp_a\vert + \vert\wp_b\vert$
a^n	$n\vert\alpha\vert(\vert A\vert)^{n-1}$	$n\vert\wp_a\vert$
$\sqrt[n]{a}$	$\dfrac{\vert\alpha\vert}{n\vert\sqrt[n]{A^{n-1}}\vert}$	$\dfrac{1}{n}\vert\wp_a\vert$

treme left of the approximate number and on the left of the apparent error of the approximate number.

5. The absolute errors and the relative errors of the numerical results of the various operations are given symbolically in Table 11. The symbolism used in this table is consistent with the symbolism that was employed in the development of the theory of Approximate Computation.

The expressions for the apparent errors of the numerical results of the operations indicated in Table 11 are similar in form and in content to the expressions for their corresponding absolute errors.

CHAPTER XIX

APPROXIMATE COMPUTATION OF THE SECOND TYPE

1. The Problem

In problems of Approximate Computation of the second type it is required to perform the numerical computations with the stated approximate numerical data in such a manner that a result with a definitely *predetermined* approximativeness will be obtained, after all the stated computations are performed. This problem can be restated as follows:

> *A set of approximate numerical data is given. It is required to perform certain numerical computations with it in such a manner that the final numerical result of the performed computations will be such that its approximativeness will not exceed a definitely predetermined limit.*

The expression "a definitely predetermined limit" signifies that the definite magnitude of an error of the final numerical result cannot be given in advance. All that can be predetermined is that the absolute (apparent) error of the final numerical result will be of a certain order in comparison with some absolute (apparent) error of the stated numerical data; that the relative error of the final numerical result will be of a certain order in comparison with some relative error of the stated numerical data; and that the final numerical result will be correct to a given number of significant digits.

It should be remembered that at no time can a result of the performed numerical operations with a given set of approximate numerical data approximate better than the least accurate or the least precise (in the case of addition or subtraction) of the approximate numerical data, with one exception, however—that of the extraction of roots, in which an extra significant digit may at times be obtained. This is the fundamental limitation in the case of problems of Approximate Computation of the second type.

The numerical operations that are to be considered are:

1. Addition.
2. Subtraction.
3. Multiplication.

4. Division.
5. Raising to a power.
6. Extraction of roots.
7. Any combination of these six operations.

It could be suggested that in the case of Approximate Computation of the second type it would be sufficient to carry out the numerical computations with the approximate data without altering the data themselves, and then the final numerical result should be altered to conform with the limitations as set forth in the particular problem. Such a procedure might be sufficient, but it is not basically necessary at all, and this will be shown in the analysis of every numerical process in detail.

In order that there may be no repetition of the analysis of the particular numerical operations we shall make use of the results that were obtained in Chapters XVI and XVII. Constant reference to these chapters will therefore be made.

We shall prove the following: If in the same approximate number, which is given as correct to k significant digits, we round off r significant digits, then the absolute error of the rounded off approximate number is of the order r in comparison with the absolute error of the approximate number in the original state, and the relative error of the rounded off approximate number is of the order r in comparison with the relative error of the approximate number in the original state.

Denote the approximate number in the original state by a. Denote the rounded off approximate number by b. Then

$$|\alpha_1| \leqslant 0.5 \cdot 10^{p+n-k+1},$$

and

$$|\beta_1| \leqslant 0.5 \cdot 10^{p+n-(k-r)+1}.$$

Then

$$\frac{|\alpha_1|}{|\beta_1|} \lessapprox 10^{-r}, \quad (r = 1, 2, 3, \cdot \cdot \cdot).$$

We have also

$$\wp'_a < \frac{1}{10^{k-1}}, \quad \text{and} \quad \wp'_b < \frac{1}{10^{k-r-1}},$$

and

$$\frac{\wp'_a}{\wp'_b} \lessapprox 10^{-r}, \quad (r = 1, 2, 3, \cdot \cdot \cdot).$$

We shall make use of this property in the analysis that follows.

II. ADDITION

Let a and b be two approximate numbers whose absolute errors are α and β respectively. Let their corresponding exact values be A and B. Let their respective relative errors be $\wp_a$ and $\wp_b$.

Assume that the approximate numbers a and b are given such that

$$\frac{|\alpha_1|}{|\beta_1|} \underset{\sim}{=} 10^{-r}, \quad (r = 0, 1, 2, 3, \cdots)$$

and it is given that

a is correct to $k + r$ significant digits,

and

b is correct to k significant digits.

It is required to obtain the sum of these two approximate numbers within a stated absolute error, given in terms of its upper limits

$$|\phi_1| > |\beta_1|.$$

Then

$$\frac{|\alpha_1|}{|\phi_1|} \underset{\sim}{=} 10^{-s}, \quad (s = 0, 1, 2, 3, \cdots),$$

and

$$\frac{|\beta_1|}{|\phi_1|} \underset{\sim}{=} 10^{-(s-r)}.$$

It follows then that both approximate numbers a and b must be rounded off a certain number of digits *before the process of* addition takes place. Since in the process of addition there may be a loss of one digit on the extreme right of the resultant sum, the approximate number a should be rounded off $(s - 1)$ digits, and the approximate number b should be rounded off $(s - r - 1)$ digits. Thus both approximate numbers a and b will be rounded off to the same place of significance. If after the addition is performed the resultant sum still has an extra digit on the extreme right, then this digit should be rounded off.

Example:

Add $(3568.7423 + 0.00004)$ and $(167.93578 + 0.00002)$ so that the sum has an apparent error not greater than 0.05 in absolute value. We have then

$$\begin{array}{r} 3{,}568.74 + 0.002 \\ 167.93 + 0.006 \\ \hline 3{,}736.67 + 0.008 \end{array}$$

or

$$3{,}736.7 \;-\; 0.02.$$

If more than two approximate numbers are to be added the same method is applicable.

Should the approximate numbers be given as correct within a stated relative error, and the approximativeness of the resultant sum be given by a predetermined upper limit of a relative error, the procedure is similar to the one illustrated above. It is necessary to determine the number of significant digits that are to be rounded off, and after all the approximate numbers are rounded off within one extra place to the right of the last place of significance of the resultant sum the addition should be performed as illustrated in the above example. Caution should be exercised in the case of addition, however, when another digit on the left may appear in the resultant sum. This happens in the case of the sum of two approximate numbers when the sum of the first significant digits on the extreme left of these numbers is greater than or equal to 10. Sometimes the first two significant digits on the extreme left result in a sum greater than or equal to 100.

III. SUBTRACTION

The procedure in subtraction is similar to that in addition. If the approximate numbers are given as correct within a stated apparent error, and the difference is to be obtained within a given upper limit of an absolute error, the approximate numbers should be rounded off as follows:

(1) If by inspection the sum of the absolute values of the apparent errors (of opposite directions) is, when the numbers are rounded off, less than $10^{p+n-k+1}$, then the approximate numbers should be rounded off to the last place of significance of the resultant difference.

(2) If by inspection it is found that the sum of the absolute values of the apparent errors (the directions being opposite) of the rounded off numbers would be equal to or greater than $10^{p+n-k+1}$, then the approximate numbers should be rounded off to one extra place to the right of the last place of significance of the resultant difference of the two approximate numbers.

The use of the relative error in the case of subtraction is decidedly disadvantageous. This is because

$$|\wp_a| < |\wp_{(a-b)}| \gtreqless |\wp_b|.$$

In many cases the resultant difference may lose several significant digits on the extreme left, and the application of the relative error test is too lengthy. It is, therefore, advisable to revert, if possible, to the absolute (apparent) error. If this is impossible, the process of subtraction should take place first, and the resultant difference should be rounded off to conform with the limitations of the problem. The same applies to the case when the approximate numbers are given as correct to a stated number of significant digits.

IV. MULTIPLICATION

In the process of multiplication the relative error plays the more prominent part. The use of the absolute (apparent) error is restricted to the final stage of the process of multiplication only, and, therefore, since the process of obtaining the absolute (apparent) error of a product by means of the formula

$$|A||\beta| + |B||\alpha|$$

is lengthy, Approximate Computation of the second type is effective only when the relative error criterion is used, or the approximate numbers are given as correct within a stated number of significant digits.

Let the two approximate numbers a and b have relative errors $\wp_a$ and $\wp_b$ respectively, and

$$\frac{\wp'_a}{\wp'_b} \eqsim 10^{r}, \quad (r = 0, 1, 2, 3, \cdot \cdot \cdot),$$

and it is required that the product (ab) have a relative error $|\wp_{(ab)} > |\wp_b|$. Let

$$\frac{\wp'_a}{\wp'_{(ab)}} \eqsim 10^{-s}, \quad (s = 1, 2, 3, \cdot \cdot \cdot),$$

then

$$\frac{\wp'_b}{\wp'_{(ab)}} \eqsim 10^{-(s-r)}.$$

From the above relations it follows that the approximate numbers may be rounded off a certain number of digits *before the process of multiplication takes place.*

[illegible] the approximate number a be correct to $(k + r)$ significant [illegible] Then the approximate number b is correct to k significant [illegible] According to the requirements of the problem the resultant product should be correct to $(k + r - s)$ significant digits $(s > r)$. The approximate number a should be rounded off, therefore, to $(k + r - s + 1)$ significant digits, and the approximate number b

should have $(s - r - 1)$ digits rounded off. If after the process of multiplication has taken place there are still $(k + r - s + 1)$ significant digits, the resultant product should be rounded off once more, and the last significant digit on the extreme right should be rounded off.

Caution should be exercised in the use of the relative error for the determination of the number of the significant digits of an approximate number. As shown,[1] the correct number of significant digits of an approximate number can be determined by means of the examination of the apparent error, the relative error, and the first significant digit on the extreme left of the approximate number. Therefore, the apparent error of the resultant product should be used to check the correctness of the performed multiplication and the consistency of this result with the limitations of the given problem.

v. DIVISION

In the case of division, which is similar to the process of multiplication as far as the properties of the relative error are concerned, if and when the relative error criterion is taken into consideration, the absolute (apparent) error is used in the final stage of the process of computation. Moreover, since the process of obtaining the apparent error by means of the formula

$$\frac{|A||\beta| + |B||\alpha|}{|B|^2}$$

is too lengthy, Approximate Computation of the second type is restricted to cases in which the approximate numbers are given as correct within a stated relative error, or they are given as correct to a stated number of significant digits, or the transfer to these two types of criteria is possible. The determination of the approximativeness of the final numerical result of the performed operation should, therefore, be stated in terms of the upper limit of the relative error or in terms of the number of significant digits.

Let two approximate numbers a and b be such that their relative errors are $\wp_a$ and $\wp_b$, and $|\wp_a| \gtreqless |\wp_b|$, so that either

$$\frac{\wp'_a}{\wp'_b} \leqq 10^{-r}, \quad \text{or} \quad \frac{\wp'_b}{\wp'_a} \leqq 10^{-r},$$
$$(r = 0, 1, 2, 3, \cdot \cdot \cdot).$$

---

[1] See Chap. XIV, sec. XIII.

Either a or b is the dividend or the divisor.[2]

It is required that the quotient $(a:b)$, or $(b:a)$ be such that its relative error

$$|\wp_{(a':b')}| > |\wp_{(a:b)}| \text{ or } |\wp_{(b':a')}| > |\wp_{(b:a)}|$$

where a' and (or) b' are the rounded off approximate numbers. Then

$$|\wp_{(a':b')}| > |\wp_b| \quad \text{or} \quad |\wp_{(b':a')}| > |\wp_b|$$

or, in the second case

$$|\wp_{(a':b')}| > |\wp_a| \quad \text{or} \quad |\wp_{(b':a')}| > |\wp_a|.$$

We shall examine the first case only. The treatment of the second case is identical with the treatment of the first case.

Let

$$\frac{\wp'_a}{\wp'_{(a':b')}} \eqsim 10^{-s}, \quad \text{or} \quad \frac{\wp'_a}{\wp'_{(b':a')}} \eqsim 10^{-s},$$
$$(s = 1, 2, 3, \cdot \cdot \cdot).$$

Then

$$\frac{\wp'_b}{\wp'_{(a':b')}} \eqsim 10^{-(s-r)} \quad \frac{\wp'_b}{\wp_{(b':a')}} \eqsim 10^{-(s-r)}.$$

From the above relations it follows that the approximate numbers may be rounded off a certain number of digits *before the process* of division takes place.

Let the approximate number a be correct to $(k + r)$ significant digits. Then the approximate number b is correct to k significant digits. According to the requirements of the problem the resultant quotient should be correct to $(k + r - s)$ significant digits $(s > r)$. The approximate number a should be rounded off, therefore, to $(k + r - s + 1)$ significant digits, and the approximate number a should have $(s - r - 1)$ digits rounded off. If after the process of division has taken place there are still $(k + r - s + 1)$ significant digits, the resultant quotient should be rounded off once more, and the last significant digit on the extreme right should be rounded off.

Caution should be exercised in the use of the relative error for the determination of the number of the significant digits of an approximate number. The correct number of significant digits of an approximate number can be determined by means of the examination of the apparent error, of the relative error, and of the first significant digit on the extreme left of the approximate number. Therefore, the apparent error of the resultant quotient should be used as a check of the correctness of the performed division and the consistency of the resultant quotient with the limitations of the given problem.

---

[2] See Chap. XVII, sec. II, C, *i*.

VI. RAISING TO A POWER

The process of raising to a power is closely related to multiplication. The use of the absolute (apparent) error in the case of the raising to a power must, therefore, be restricted to the final stage of this process. Moreover, the calculation of the apparent error of the nth power ($n = 2, 3, 4, \ldots$) of an approximate number requires much work. This can be seen from the formula for the absolute error of the nth power of the approximate number a:

$$n|\alpha|(|A|)^{n-1}.$$

Approximate Computation of the second type, therefore, is effective only when the approximate number is given as correct within a stated relative error, or when it is given as correct to a stated number of significant digits. The determination of the approximativeness of the final result of the performed operation should, therefore, be stated in terms of the upper limit of the relative error or in terms of the number of significant digits. Should the approximate number be given as correct within a stated apparent error, and the determination of the approximativeness be also stated for the final numerical result in terms of the upper limit of the absolute error, the proper procedure is to effect, if possible, a transfer to the relative error criterion, and then the Approximate Computation of the second type can be effected.

Let the approximate number a have a relative error $\wp_a$. Let the predetermined relative error of the nth power of the approximate number[3] a' be $|\wp_{(a'^n)}| > |\wp_a|$. Then

$$\frac{\wp'_a}{\wp_{(a'^n)}} \eqsim 10^{-r}, \quad (r = 1, 2, 3, \cdot \cdot \cdot).$$

The above relation shows that if $n \leqslant 9$, $(r - 1)$ significant digits of the approximate number a should be rounded off *before the process of raising to a power takes place.*[4]

Let the approximate number a be correct to k significant digits. Then this approximate number should be rounded off to $(k - r + 1)$ significant digits. If after the process of raising to a power the resultant nth power ($n \leqslant 9$) still has $(k - r + 1)$ digits, the last significant digit on the extreme right of the resultant nth power ($n \leqslant 9$) should be rounded off.

---

[3] We denote by a' the rounded off approximate number a.

[4] See Chap. XVI, sec. IV, C, *iv*.

If $10 \leqslant n \leqslant 99$, then the approximate number a should be rounded off to $(k - r + 2)$ significant digits, and so on.

Caution should be exercised in the use of the relative error for the determination of the number of the significant digits of the approximate number representing the nth power of the approximate number a'. The correct number of significant digits of an approximate number can be determined by means of the examination of the apparent error, of the relative error, and of the first significant digit on the extreme left of the approximate number. Therefore, the apparent error of the resultant nth power should be used to check the correctness of the performed raising to a power and the consistency of the resultant nth power with the limitations of the given problem.

VII. Extraction of Roots

In the process of extraction of roots, as in the processes of multiplication, division, and raising to a power, Approximate Computation of the second type must be restricted to determining the approximativeness of the final numerical result by the relative error or by the number of the significant digits. The use of the absolute (apparent) error is disadvantageous, as can be seen from the formula for the absolute error of the nth root of an approximate number a:

$$\frac{|\alpha|}{n\left|\sqrt[n]{A^{n-1}}\right|}.$$

Should the approximate number a be given as correct within a stated apparent error, and the determination of the approximativeness of the final numerical result be also stated in terms of an upper limit of the absolute error, the proper procedure is to effect, if possible, a transfer to the relative error criterion, and then the Approximate Computation of the second type can be effected.

Let the approximate number a have a relative error $\wp_a$. Let the predetermined relative error of the nth root of the approximate number[5] a' be $|\wp_{\sqrt[n]{a'}}| > |\wp_a|$. Then

$$\frac{\wp'_a}{\wp'_{\sqrt[n]{a'}}} \eqsim 10^{-r}, \quad (r = 1, 2, 3, \cdot \cdot \cdot).$$

The above relation shows that the approximate number a should be rounded off a certain number of digits *before the process of extraction of roots takes place.*

---

[5] We denote by a' the rounded off approximate number a.

Let the approximate number a be correct to k significant digits. Then this approximate number should be rounded off, if $n \leqslant 9$, to $(k - r)$ significant digits. If after the process of extraction of the root the resultant nth root $(n \leqslant 9)$ has $(k - r + 1)$ significant digits,[6] the last significant digit on the extreme right of the resultant nth root $(n \leqslant 9)$ should be rounded off.

If $10 \leqslant n \leqslant 99$, then the approximate number a should be rounded off to $(k - r - 1)$ significant digits. This can be readily seen from the expression for the relative error of the nth root of the approximate number a:

$$|\wp_{\sqrt{a}}| \doteq \frac{1}{n}\,|\wp_a|.$$

Caution should be exercised in the use of the relative error for the determination of the number of the significant digits of the approximate number representing the nth root of the approximate number a'. The correct number of significant digits of an approximate number can be determined by means of the examination of the apparent error, of the relative error, and of the first significant digit on the extreme left of the approximate number. Therefore, the apparent error of the resultant nth root should be used as a check of the correctness of the performed extraction of the root and the consistency of the resultant nth root with the limitations of the given problem.

VIII. COMPLEX APPROXIMATE COMPUTATION

The various combinations of the above six numerical operations have been analyzed in Chapter XVIII.

The most effective method in Approximate Computation of the second type is the use of the relative error criterion, or of the significant digits criterion. In all the operations, except addition and subtraction, the use of the absolute (apparent) error renders Approximate Computation of the second type almost impossible because of the amount of work it involves. It is therefore advisable to transfer, whenever possible, to the relative error criterion for the approximativeness of the numerical data. If this cannot be done, the work with the numerical data should be carried to the point where this transfer can be made.

---

[6] See Chap. XVII, sec. III, C.

IX. THE DETERMINATION OF THE DATA WHOSE APPROXIMATIVENESS IS GIVEN IN ADVANCE

Very often it is required to obtain numerical data in such a manner that a numerical result of a predetermined approximativeness is obtained. No general rules in such a case can be given. But some methods of procedure may be outlined, and these methods are common to all cases.

(1) If a numerical result is to be obtained, and its approximativeness is given in advance, the computer should determine the type of errors he will use, that is, whether the use of the absolute (apparent) errors, or the use of the relative errors will be more effective. As a rule, for all processes except addition and subtraction relative errors are more effective.

(2) Data that may be obtained from tables should be assigned the least possible errors (relative), so that the weight of the errors falls on data that are to be obtained by means of measurement.

(3) To all data that are to be obtained by means of measurement an equal amount of errors should be assigned, unless it is possible to perform measurements and secure data with errors of lower orders.[7]

Examples:

1. It is required to compute the volume of a cylinder

$$V = \pi r^2 h,$$

with an error not exceeding 400 cm³, if

$$r = (25.37 \pm 0.003) \text{ cm.},$$

and

$$h = (36.15 \pm 0.004) \text{ cm.}$$

Roughly, we have $3 \cdot 25^2 \cdot 36 = 67{,}500$.

Then

$$\wp_V = \frac{400}{67{,}500} = 0.006.$$

In other words, the relative error of the volume is 0.006. We can assign to π the smallest relative error possible. Assigning to it 0.001, or even 0.0005, we determine the number of significant digits in π. Then we split the remaining portion of the relative error, that is, 0.005 or 0.0055, among r^2 and h; that is, the remaining portion must be divided into three parts, and two of them assigned to r^2 and one to h. Thus we determine the number of significant digits in r and h.

---

[7] See Chap. XIV, sec. XVI.

2. Obtain the perimeter of a rectangle whose sides are $a = (425.67 \pm 0.003)$ ft., and $b = (342.84 \pm 0.005)$ ft., so that the error of the perimeter does not exceed 1 foot. Roughly, the perimeter is 1,534 ft. Then

$$1 \approx 850\,|\wp_a| + 684\,|\wp_b|.$$

We shall assume that both sides have been measured with the same precision, and since the approximate numbers have the same number of significant digits, we shall assume that the accuracies of the measures are of the same order, that is,

$$\frac{\wp'_a}{\wp'_b} \eqsim 1.$$

Then

$$1 \approx 1{,}534\,\wp'_a \approx 1{,}534\,\wp'_b,$$

or

$$\wp'_a \approx \wp'_b \approx 0.00066.$$

Estimating the first significant digits on the left of the respective absolute (apparent) errors we get

$$\alpha \approx \pm 0.3, \quad (\alpha \approx \pm 0.28)$$

and

$$\beta \approx \pm 0.2, \quad (\beta \approx \pm 0.22).$$

Then the perimeter is

$$2 \cdot [(425.7 \pm 0.3) + (342.9 \pm 0.2)]$$

or

$$1{,}537 \pm 1.$$

Rounding off we have $1{,}540 \pm 4$.
This shows that we exceeded the limit of the error. Hence, we shall take the lengths of the sides correct to 4 significant digits each. We have then

$$a \approx 425.7 \pm 0.03,$$

and

$$b \approx 342.8 \pm 0.05.$$

Then, the perimeter is

$$2[(425.7 \pm 0.03) + (342.8 \pm 0.05)]$$

or

$$1{,}537. \pm 0.2.$$

3. What should be the approximativeness of the side of a cube, so that the error of its volume does not exceed k cm^3?

We have then, letting the side of the cube be $a = A \pm \alpha$, and $|\wp_a| = \frac{|\alpha|}{|A|}$. Then the relative error of the volume is $3|\wp_a|$, and

$$k = 3|\wp_a|A^3 = 3|\alpha|A^2.$$

Then

$$|\alpha| \approx \frac{k}{3A^2}.$$

This formula will give numerical values after the values of k and A^2 are substituted in it.

CHAPTER XX

LOGARITHMS

I. THE METHOD OF PROPORTIONAL PARTS

IN APPROXIMATE Computation considerable use is made of tables of logarithms, most commonly the 4-, 5-, and 7-place tables. The tables of logarithms are so constructed that the calculation of logarithms of numbers not found in the table (or the reverse, i.e., the determination of a number not found in the table, but whose logarithm is given) can be effected by means of the first differences of the two neighboring values found in the table, one in defect, and the other in excess. This method is known as the Method of Proportional Parts.

If the logarithms are taken to the base e[1] (these are known as the natural logarithms) we have the following series for $\log (1 + x)$ to the base e, when $|x| < 1$.

$$\log_e(1 + x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdot\cdot\cdot - (-1)^n\frac{x^n}{n} + \cdot\cdot\cdot .$$

The transformation of the natural logarithms to the logarithms at the base 10, known as the Briggsian logarithms, is effected by means of the formula

$$\log_{10}N = M \log_e N,$$

where M is known as the "modulus."[2]

Let N and $N_1 = N + n$, $(|n| < |N|)$ be two numbers. Then

$$d = \log_e(N + n) - \log_e N = \log_e \frac{N + n}{N} = \log_e\left[1 + \frac{n}{N}\right].$$

---

[1] e is defined as follows:

$$e = \underset{n \to \infty}{\text{Limit}} \left[1 + \frac{1}{n}\right]^n.$$

e is a transcendental number. Its numerical value, correct to 16 significant digits is

$$e = 2.71828 \quad 18284 \quad 39045 + 3 \cdot 10^{-16}$$

[2] The "modulus" is a transcendental number. The most recently computed value of the modulus was given by Adams as correct to 272 (probably 273) significant digits; (Proceedings of the Royal Society, 1886, Vol. 42, p. 22). The numerical value of the modulus correct to the first 35 significant digits is

$$0.43429 \quad 44819 \quad 03251 \quad 82765 \quad 11289 \quad 18916 \quad 60508 + 3 \cdot 10^{-36}.$$

Hereafter, we shall write the logarithms to the base 10 without the subscripts. Then

$$\log\left[1+\frac{n}{N}\right]=M\left\{\frac{n}{N}-\frac{1}{2}\left[\frac{n}{N}\right]^2+\frac{1}{3}\left[\frac{n}{N}\right]^3-\cdots\right.$$
$$\left.\cdots-\frac{(-1)^k}{k}\left[\frac{n}{N}\right]^k+\cdots\right\},$$

where M is the modulus.

The magnitude of the ratio $\frac{n}{N}$ is chosen such that the terms of the series

$$-\frac{1}{2}\left[\frac{n}{N}\right]^2, +\frac{1}{3}\left[\frac{n}{N}\right]^3, \text{ etc.}$$

are so small in comparison with the ratio $\frac{n}{N}$ itself, that their numerical values do not affect at all the last digit on the extreme right of the number representing the logarithm which is found in the table of logarithms. This is the underlying principle of the Method of Proportional Parts.

The difference between the logarithms of two numbers, N and $(N+1)$, is in every table of logarithms for that table

$$d=M\frac{1}{N}.$$

Suppose that we have a number x such that

$$N<x<(N+1).$$

Assume that

$$x=N+n,$$

then

$$\log x=\log N+M\frac{n}{N}=\log N+nd.$$

This is the Method of Proportional Parts.[3]

---

[3] This Method of Proportional Parts corresponds to the process of interpolation to the first two terms of the Newton-Gregory interpolation formula (Chapter XIV, sec. III, C). There are methods of interpolation by means of which more precise numerical values of the logarithms can be obtained. These are the Newton-Gauss interpolation formulas, the central difference interpolation formulas (the Everett formula, the Bessel formula, and the Stirling formula).

See Whittaker, E. T. and Robinson, G.: The Calculus of Observations; Blackie and Son, Ltd., London, 1926; Chaps. I and III. Steffensen, J. F.: Interpolation;

II. THE PRECISION OF TABLES OF LOGARITHMS

The magnitude of the ratio $\frac{n}{N}$ as shown in the preceding pages must be such that the terms of the series for the log $\left[1 + \frac{n}{N}\right]$ containing the second, the third, and the higher powers of this ratio do not affect the last significant digit on the right of the logarithm.

If we take

$$\frac{n}{N} < \frac{1}{100},$$

then

$$\frac{1}{2}\left[\frac{n}{N}\right]^2 < \frac{1}{20,000} = 0.00005.$$

This does not affect the fourth (the last on the extreme right) significant digit. Thus, in the 4-place tables of logarithms the arguments (numbers) are taken from 100 to 1000 in intervals (steps) of 1.

If we take

$$\frac{n}{N} < \frac{1}{1,000},$$

then

$$\frac{1}{2}\left[\frac{n}{N}\right]^2 < \frac{1}{2,000,000} = 0.0000005.$$

This does not affect the fifth place or significant digit (the last on the extreme right). Thus, in 5-place tables of logarithms the arguments (the numbers) are listed from 1,000 to 10,000 in intervals (steps) of 1.

If we take

$$\frac{n}{N} < \frac{1}{10,000},$$

then

$$\frac{1}{2}\left[\frac{n}{N}\right]^2 < \frac{1}{200,000,000} = 0.000000005.$$

The Williams & Wilkins Co., Baltimore, 1927; sec. 4; pp. 22–33. Newcomb, S.: Logarithmic and Other Mathematical Tables: Holt, 1882; pp. 56–72.

Václav Hruskă: Lineárni interpolace v logarithmických tabulkách. Časopis pro Pestováni Matematiky a Fysiky, Vol. 64, no. 1. Praha, 1934, pp. R 1, ff.

On the *precision* of interpolation see: Whittaker and Robinson: *op. cit.*; there is given on p. 49 a list of works by W. F. Sheppard, D. C. Fraser, and G. J. Lidstone. Also: Lehman, Inge: On the Accuracy of Interpolation, Transactions of the Sixth Scandinavian Mathematical Congress, Copenhagen, 1926, p. 375.

This does not affect the seventh (the last on the extreme right) significant digit. Thus, in the case of the 7-place tables of logarithms the arguments (the numbers) are listed from 10,000 to 100,000 in intervals (steps) of 1.

The logarithms which are given in the tables have apparent errors, the corresponding absolute errors of which have an upper limit α_1 such that

$$|\alpha_1| \leqslant 0.5 \cdot 10^{-k},$$

k being the number of the digits in the mantissa of the logarithm. It follows then that even if a number is exact (except when it is an integral power of 10), its logarithm is always approximate.

Let the approximate number a be such that

$$|a| \leqslant |A| + |\alpha|.$$

Then

$$\log |a| \leqslant \log (|A| + |\alpha|).$$

But

$$\log (|A| + |\alpha|) = \log \left\{(|A|)\left[1 + \frac{|\alpha|}{|A|}\right]\right\} = \log (|A|) + \log (1 + |\wp_a|).$$

Hence,

$$\log |a| \leqslant \log (|A|) + M|\wp_a|.$$

Thus, $M|\wp_a|$ is the absolute error (apparent error) of $\log a$. But since,

$$M\wp'_a \leqslant 0.5 \cdot 10^{-k},$$

k being the number of the digits in the mantissa of the logarithm, and since $M < 0.5$, we have:

$$\wp'_a < 10^{-(k-1)}.$$

In other words, if the approximate number a is correct to k significant digits, then the mantissa of its logarithm must have k or $(k - 1)$ digits. This shows that it would be absurd to use a 7-place table of logarithms in computations with 3- or 4-place numbers. Conversely, a 4-place table should never be used for the calculation of logarithms of numbers containing more than five significant digits.

If a logarithm of an approximate number a is given such that

$$\log |a| \leqslant \log (|A|) + M|\wp_a|,$$

where

$$M\wp'_a \leqslant 0.5 \cdot 10^{-k},$$

then

$$\wp'_a < 10^{-k}.$$

It should be remembered, however, that the relative error test alone, without the use of the absolute (apparent) error and of the first significant digit on the extreme left of the approximate number, is ineffective. If the first significant digit of the approximate number is $x_1 \geqslant 5$, and remembering that the upper limit of the absolute error of the approximate number is

$$|\alpha_1| \leqslant 0.5 \cdot 10^{p+n-k+1},$$

k being the number of the significant digits of the approximate number, and if

$$\wp'_a < 10^{-(t-1)},$$

then the approximate number is correct to $(t - 1)$ significant digits. Hence, we conclude that when $x_1 \geqslant 5$, and the approximate number is correct to k significant digits, the mantissa of its logarithm should have k digits. The same holds true when the first significant digit of the apparent error of the approximate number a is equal to or is less than the first significant digit x_1 on the extreme left of the approximate number a.[4]

It is therefore necessary to use the relative error as well as the apparent error while working with logarithms of approximate numbers.

III. OPERATIONS WITH LOGARITHMS

The numerical operations that can be performed with logarithms are: multiplication, division, raising to a power, and extraction of roots.[5]

Since the logarithms themselves are approximate numbers, the performance of a numerical process with logarithms is subject to the rules of the operations with approximate numbers. Since the logarithms which are obtained from tables are approximate numbers, *every logarithm should and must be stated with its respective apparent error.*

A. MULTIPLICATION

Let a and b be two approximate numbers, whose absolute errors are α and β respectively, and whose relative errors are $\wp_a$ and $\wp_b$

---

[4] See Chap. XIV, sec. XIII.

[5] Addition and subtraction can be performed by means of the Gaussian Logarithms. See Zech: Taffeln der Additions und Subtraktions Logarithmen fur sieben Stellen; Berlin, 1863. Also, Newcomb, S.: *op. cit.*

respectively. Let the exact values of the approximate numbers a and b be A and B respectively such that

$$0 < A_2 < A < A_1,$$

and

$$0 < B_2 < B < B_1.$$

Then

$$|a| \leqslant |A| + |\alpha|,$$

and

$$|b| \leqslant |B| + |\beta|.$$

We have then

$$\log |a| \leqslant \log (|A|) + M|\wp_a|,$$

and

$$\log |b| \leqslant \log (|B|) + M|\wp_b|.$$

Then

$$\log |ab| \leqslant \log (|A|) + \log (|B|) + M(|\wp_a| + |\wp_b|).$$

Since

$$\alpha + \epsilon_1 = \alpha_1, \quad (\epsilon_1 \rightarrow 0),$$

and

$$\beta + \epsilon_2 = \beta_1, \quad (\epsilon_2 \rightarrow 0),$$

then

$$\operatorname*{Limit}_{\substack{\epsilon_1 \to 0 \\ \epsilon_2 \to 0}} M(|\wp_a| + |\wp_b|) \leqslant M(\wp'_a + \wp'_b).$$

The above shows that there is a very close relation between the relative error of an approximate number and the absolute (apparent) error of the logarithm of the same approximate number.

We have then: The absolute error of the logarithm of the product of two approximate numbers a and b, whose relative errors are $\wp_a$ and $\wp_b$ respectively, may equal in absolute value the sum of the absolute values of the absolute errors of the logarithms of the approximate numbers a and b respectively. The expression for the apparent error of the logarithm of the product of two approximate numbers is similar in form and content to the expression for the absolute error of the logarithm of the product of the two approximate numbers. The upper limit of the absolute error of the logarithm of the product of two approximate numbers a and b, whose respective relative errors are $\wp_a$ and $\wp_b$, the upper limits of these relative errors being $\wp'_a$ and $\wp'_b$ respectively, is equal to or is less than the sum of the upper limits of the absolute errors of the logarithms of the approximate numbers a and b respectively.

The above can be extended to the general case of the product of n approximate numbers a, b, c, etc. We have then

$$\log(|abc \cdot \cdot \cdot n|) \leqslant \sum_{K=A}^{N} \log(|K|) + M\sum_{k=a}^{n}|\wp_k|,$$

and

$$\underset{\substack{\epsilon_i \to 0 \\ (i = 1, 2, 3, \ldots, n)}}{\text{Limit}} M\sum_{k=a}^{n}|\wp_k| \leqslant M\sum_{k=a}^{n}\wp'_k$$

From the above we have that the product of (at the most), nine approximate numbers will have a logarithm such that the absolute (apparent) error of the product will be of the order "zero" or of the order "one" in comparison with the greatest (in absolute value) absolute (apparent) error of the logarithms of one of the approximate numbers which enter as factors into the product. The product of at least ten and not more than 99 approximate numbers will have a logarithm such that the absolute (apparent) error of the logarithm of the product will be of the order "one" or of the order "two" in comparison with the greatest (in absolute value) absolute (apparent) error of the logarithms of one of the numbers which enter as factors into the product.[6]

The rules concerning the directions of the absolute (apparent) errors of the factors, and the directions of the absolute (apparent) error of the logarithm of the product of the approximate numbers are the same as the rules for the directions of the absolute (apparent) errors and of the relative errors of the product of the approximate numbers. The reader is referred to Chapter XVI, sec. IV for a detailed analysis of these rules and for a statement of the correct procedure.

B. DIVISION

Let a and b be two approximate numbers, whose absolute errors are α and β respectively, and whose relative errors are $\wp_a$ and $\wp_b$ respectively. Let the exact values of the two approximate numbers a and b be A and B respectively such that

$$0 < A_2 < A < A_1,$$

and

$$0 < B_2 < B < B_1.$$

Then

$$|a| \leqslant |A| + |\alpha|,$$

and

$$|b| \leqslant |B| + |\beta|.$$

---

[6] See Chap. XVI, sec. IV, a, *iv*.

We have then

$$\log |a| \leqslant \log (|A|) + M|\wp_a|,$$

and

$$\log |b| \leqslant \log (|B|) + M|\wp_b|.$$

(1) Consider the case of $\frac{a}{b}$.

Then

$$\log (|a : b|) \leqslant \log (|A|) - \log (|B|) + M(|\wp_a| + |\wp_b|).$$

Since

$$\alpha + \epsilon_1 = \alpha_1, \quad (\epsilon_1 \rightarrow 0),$$

and

$$\beta + \epsilon_2 = \beta_1, \quad (\epsilon_2 \rightarrow 0),$$

$$\operatorname*{Limit}_{\substack{\epsilon_1 \to 0 \\ \epsilon_2 \to 0}} M(|\wp_a| + |\wp_b|) \leqslant M(\wp'_a + \wp'_b).$$

(2) Consider the case of $\frac{b}{a}$.

Then

$$\log (|b : a|) \leqslant \log (|B|) - \log (|A|) + M(|\wp_a| + |\wp_b|).$$

Since

$$\alpha + \epsilon_1 = \alpha_1, \quad (\epsilon_1 \rightarrow 0),$$

and

$$\beta + \epsilon_2 = \beta_1, \quad (\epsilon_2 \rightarrow 0),$$

$$\operatorname*{Limit}_{\substack{\epsilon_1 \to 0 \\ \epsilon_2 \to 0}} M(|\wp_a| + |\wp_b|) \leqslant M(\wp'_a + \wp'_b).$$

We have then: The absolute error of the logarithm of the quotient of two approximate numbers a and b, whose relative errors are $\wp_a$ and $\wp_b$ respectively, may equal in its absolute value the sum of the absolute values of the absolute errors of the logarithms of the approximate numbers a and b respectively.

The expression for the apparent error of the logarithm of the quotient of the two approximate numbers a and b is similar in its form and content to the expression for the absolute error of the logarithm of the quotient of these two approximate numbers. The upper limit of the absolute error of the logarithm of the quotient of two approximate numbers a and b, whose relative errors are $\wp_a$ and $\wp_b$ respectively, the upper limits of these relative errors being $\wp'_a$ and $\wp'_b$ respectively, is equal to or is less than the sum of the upper limits of the absolute errors of the logarithms of the approximate numbers a and b respectively.

The rules concerning the directions of the absolute (apparent) errors of the logarithms of the dividend and of the divisor and the direction of the absolute (apparent) error of the logarithm of the quotient of two approximate numbers are the same as the rules for the directions of the absolute (apparent) errors and of the relative errors of the quotient of two approximate numbers. The reader is referred to Chapter XVII, sec. II for a detailed analysis of these rules and for a statement of the correct procedure.

The analysis of the accuracy of the quotient of two approximate numbers a and b was developed in detail in pages 170–171 and pages 174–177 and the precision of the logarithm as well as the accuracy of the logarithm of the quotient of two approximate numbers should conform with the results of Chapter XVII, sec. II. Moreover, since the absolute (apparent) error of the logarithm of the quotient of two approximate numbers is actually the absolute (apparent) error of the difference of two approximate numbers (the logarithms are approximate numbers), the results concerning the absolute (apparent) error of a difference of two approximate numbers should be kept in mind.

The analysis of the approximativeness of the quotient of two approximate numbers should be effected by means of the absolute (apparent) and the relative error and the results of Chapter XVII, sec. II.

C. RAISING TO A POWER

Let a be an approximate number, whose absolute error is α, and whose relative error is $\wp_a$. Let the exact value of the approximate number a be A such that

$$0 < A_2 < A < A_1.$$

Then

$$|a| \leqslant |A| + |\alpha|,$$

$$\log(|a^n|) \leqslant n \log(|A|) + nM|\wp_a|.$$

Since

$$\alpha + \epsilon_1 = \alpha_1, \quad (\epsilon_1 \rightarrow 0),$$

$$\operatorname*{Limit}_{\epsilon_1 \to 0} nM|\wp_a| \leqslant nM\wp'_a.$$

We have then: The absolute error of the logarithm of the nth power of the approximate number a, whose relative error is $\wp_a$, is equal in its absolute value to n times the absolute value of the absolute error of the logarithm of the approximate number a itself. The expression for the apparent error of the logarithm of the nth power of the approximate number a is similar in form and content to the expression for the absolute error of the logarithm of the *n*th power of the approximate number a. The upper limit of the absolute error of the logarithm of the *n*th power of the approximate number a, whose relative error is $\wp_a$, the upper limit of the relative error being $\wp'_a$, is equal to or is less than n times the upper limit of the absolute error of the logarithm of the approximate number a itself.

The logarithm of at the most the ninth power of an approximate number will have an absolute (apparent) error which will be of the order "zero" or of the order "one" in comparison with the absolute (apparent) error of the logarithm of the approximate number itself. The logarithm of at the least the tenth power and at the most the 99th power of an approximate number will have an absolute (apparent) error which will be of the order "one" or of the order "two" in comparison with the absolute (apparent) error of the logarithm.[7]

The rules concerning the direction of the absolute (apparent) error of the logarithm of the *n*th power of the approximate number a are the same as the rules for the direction of the absolute (apparent) error and the relative error of the *n*th power of the approximate number itself. The direction of the absolute (apparent) error of the logarithm of the *n*th power of the approximate number a is the same as the direction of the absolute (apparent) error of the logarithm of the approximate number a itself, and the direction of the absolute (apparent) error of the logarithm of the approximate number a is the same as the direction of the absolute (apparent) error of the approximate number a itself.

D. EXTRACTION OF ROOTS

Let a be an approximate number, whose absolute error is α, and whose relative error is $\wp_a$. Let the exact value of the approximate number a be such that

$$0 < A_2 < A < A_1.$$

Then

$$|a| \leqslant |A| + |\alpha|,$$

$$\log\left(|\sqrt[n]{a}|\right) \leqslant \frac{1}{n}\log\left(|A|\right) + \frac{1}{n}M|\wp_a|.$$

Since

$$\alpha + \epsilon_1 = \alpha_1, \quad (\epsilon_1 \rightarrow 0),$$

$$\operatorname*{Limit}_{\epsilon_1 \rightarrow 0} \frac{1}{n}M|\wp_a| \leqslant \frac{1}{n}M\wp'_a.$$

---

[7] See Chap. XVI, sec. IV, C, *iv*.

We have then: The absolute error of the logarithm of the nth root of the approximate number a, whose relative error is $\wp_a$, is equal in its absolute value to $\frac{1}{n}$ of the absolute error of the logarithm of the approximate number a itself. The expression for the apparent error of the logarithm of the nth root of the approximate number a is similar in form and content to the expression for the absolute error of the logarithm of the nth root of the approximate number a. The upper limit of the absolute error of the logarithm of the nth root of the approximate number a is equal to or is less than $\frac{1}{n}$ of the upper limit of the absolute error of the logarithm of the approximate number a.

The logarithm of at the most the ninth root of the approximate number a will have an absolute error (apparent error) which will be such that the absolute (apparent) error of the logarithm of the approximate number a will be of the order "zero" or of the order "one" in comparison with the absolute (apparent) error of the logarithm of the root of the approximate number.

The logarithm of at the least the tenth root and of at the most the 99th root of the approximate number a will have an absolute (apparent) error such that the absolute (apparent) error of the logarithm of the approximate number a will be of the order "one" or of the order "two" in comparison with the absolute (apparent) error of the logarithm of the root of the approximate number.

The rules for the direction of the absolute (apparent) error of the logarithm of the nth root of the approximate number a are the same as the rules for the direction of the absolute (apparent) error and the relative error of the nth root of the approximate number itself. The direction of the absolute (apparent) error of the logarithm of the nth root of the approximate number a is the same as the direction of the absolute (apparent) error of the logarithm of the approximate number a itself, and this direction is the same as the direction of the absolute (apparent) error of the approximate number a itself.

The above analysis shows that in the case of extraction of roots very often a given table of logarithms may be insufficiently precise for the purpose of calculating the roots. Additional significant digits may be obtained in roots of approximate numbers. In such cases it is necessary to calculate the root (after its logarithm is obtained) by means of a table of logarithms with a corresponding number of places in the mantissas.

IV. THE ABSOLUTE ERROR OF THE LOGARITHM OF AN APPROXIMATE NUMBER

Let a be an approximate number, whose absolute error is α, and whose relative error is $\wp_a$. Let the exact value of the approximate number be A, such that

$$0 < A_2 < A < A_1.$$

Then

$$|a| \leqslant |A| + |\alpha|,$$

and

$$\log |a| \leqslant \log (|A|) + M|\wp_a|.$$

In the analysis of the four operations which can be performed by means of logarithms (see sections II and III of this chapter) the expression $\log (|A|)$ was used without any reference to the $\log (|A|)$ itself. Since the logarithms of all numbers, be they exact or approximate (with one exception in the case of exact numbers, when these numbers are of the form 10^m, m being an integer), are all approximate, the $\log (|A|)$ must have an absolute error of its own.

Let then

$$\log (|A|) \leqslant |S| + |\sigma|,$$

then

$$\log |a| \leqslant |S| + |\sigma| + M|\wp_a|.$$

If the approximate number a is given as correct to k significant digits, then the mantissa of $\log (|A|)$, whose exact value is S, contains either k or $(k - 1)$ digits (see section II of this chapter). The absolute error of $\log |a|$ is then

$$|\sigma| + M|\wp_a|.$$

Should $|\sigma| + M|\wp_a|$ be greater than 10^{-k}, then in order that k digits of the mantissa of the logarithm of the approximate number be preserved, the absolute (apparent) error of the $\log |a|$ is represented by the units part of

$$|\sigma| + M|\wp_a|,$$

so that the absolute (apparent) error of the logarithm of the approximate number is still represented in the form

$$\leqslant 0.5 \cdot 10^{-k},$$

and the excess of $|\sigma| + M|\wp_a|$ over this value of the absolute (apparent) error becomes a part of the true value of the logarithm of the approximate number *a*. For example: Obtain the log (3.472 + 0.0003). We have then

$$
\begin{array}{rl}
\log 3.470 & = 0.5403 + 0.00003 \\
2 & = 0.0002 + 0.00005 \\
M|\varphi_a| = 0.0003 & = \qquad\qquad + 0.00004 \\
\hline
\log(3.472 + 0.0003) & = 0.5406 + 0.00002
\end{array}
$$

Tables of logarithms should be constructed as follows:

1. A k-place table of logarithms should have mantissas given to k places and to $(k + 1)$ places. The descriptive word "place" should refer not to the number of the decimal places of the mantissa of the logarithm in this case, but to the number of the significant digits to which the approximate number is correct. The $(k + 1)$-place table of logarithms should be used for approximate numbers which are given as correct to $(k + 1)$ significant digits, but whose apparent errors have their first significant digits equal to or less than the first significant digits on the extreme left of the approximate numbers to which these apparent errors correspond.

2. To the right of every mantissa there should be in the table of logarithms a column with the apparent errors of the logarithms (the mantissas represent the true values) of the numbers.

3. Since it is inconvenient to tabulate apparent errors of positive and negative directions, all the apparent errors of the logarithms given in the tables (these are the logarithms of the true values of the approximate numbers) should be tabulated as positive, and on account of this, the apparent errors of the logarithms of the true values are of the form

$$10^{-k} > \text{apparent error} \geqq 10^{-(k+1)},$$

and some apparent errors may be 0 (zero), but this does not signify that the logarithm is an exact number.

4. If the proportional parts, or if the "mean differences" are tabulated, the apparent error portion of the proportional parts or of the mean differences should be tabulated alongside, on the right, of each individual proportional part or mean difference.

Since all apparent errors of the logarithms are tabulated as positive, should an approximate number whose logarithm is to be calculated have a negative apparent error, the upper limit of the true value of the approximate number should be replaced by the lower limit of the true value of the approximate number. This will change the direction of the apparent error of the approximate number *a* from negative to positive.

All other rules for computation with and calculation of the logarithms given in the textbooks of algebra hold good for the logarithms given in the tables of the type described above.

The expressions "four-place" tables, "five-place" tables, etc., should refer rather to the number of the significant digits in the approximate number than to the number of the digits in the mantissa of a logarithm.

CHAPTER XXI

TRIGONOMETRIC FUNCTIONS

1. The Method of Proportional Parts

The trigonometric functions which find place in the numerical problems of mathematics are tabulated in tables of two kinds.

(1) The trigonometric functions, which are ratios, are tabulated in their numerical values. The tables of these values are known as the tables of the "natural" trigonometric functions (natural sines, natural cosines, natural tangents, etc.) of angles.

(2) The logarithms of the numerical values of the trigonometric functions are given to a certain number of places, digits in the mantissas. These tabulations are known as the tables of the "logarithmic trigonometric functions" (logarithmic sines, logarithmic cosines, logarithmic tangents, etc.).

The table of the trigonometric functions is so constructed that the calculation of the numerical values of the trigonometric functions or of the logarithmic trigonometric functions not found in the table (or the converse, i.e., the determination of the angle whose trigonometric functions, natural or logarithmic, are given, but not found in the table) can be effected by means of the first differences of the two neighboring values found in the table, one in defect and the other in excess. This method is based on the principle: "*except for angles which are either very small or very nearly equal to a right angle, small changes in the natural or in the logarithmic functions of an angle are proportional to the change in the angle itself.*"[1]

In the following analysis we shall use the circular measure for angles. We shall denote a trigonometric function by $f(x)$, be it logarithmic or natural. Then, by means of Taylor's theorem, we have

$$f(x+h) = f(x) + hf'(x) + \tfrac{1}{2}h^2f''(x+\theta h),$$

where[2] $0 < \theta < 1$, and $\frac{1}{2}h^2f''(x+\theta h)$ is the error made, if we assume that

---

[1] Hobson, E. W.: A Treatise on Plane Trigonometry; Cambridge, at the University Press, 1921; p. 146.

[2] The following is based on the treatment of this problem of the principle of Proportional Parts by Hobson: *op. cit.*

$$f(x+h) - f(x) = hf'(x),$$

where

$$\operatorname*{Limit}_{h \to 0} \frac{f(x+h) - f(x)}{h} = f'(x).$$

If $\frac{1}{2}h^2f''(x+\theta h)$ is so small that it does not affect the last significant digit on the extreme right of the numerical value of the trigonometric function, the proportionality between the increase of the angle and the increase of the function is established. There are two exceptions, however:

(1) If the angle x is such that $f'(x)$ is very small, then the difference $f(x+h) - f(x)$ may tend to zero. In such a case two or more tabulated values of $f(x)$ may be the same.

(2) If the angle x is such that $\frac{1}{2}h^2f''(x+\theta h)$ is large in comparison with $hf'(x)$ itself, then the difference $f(x+h) - f(x)$ is not proportional to the increase of the angle h.

In the first case (1), the differences

$$f(x+h) - f(x)$$

are *insensible*. In the second case (2), the differences

$$f(x+h) - f(x)$$

are *irregular*. In both cases (1) and (2) the Method of Proportional Parts fails. There are special methods, however, by means of which these difficulties may be overcome.[3] We shall not concern ourselves with these special methods since they have been fully developed in standard treatises on trigonometry.

If $f(x) = \sin x$,
then

$$f(x+h) - f(x) = h\cos x - \tfrac{1}{2}h^2\sin(x+\theta h).$$

When x is nearly a right angle, difference

$$\sin(x+h) - \sin x$$

is insensible, because $h\cos x$ is very small. This difference is then irregular also, because $\frac{1}{2}h^2\sin(x+\theta h)$ may become comparable in magnitude with $h\cos x$.

---

[3] See Hobson: *op. cit.*, pp. 150-154. Chauvenet, W.: A Treatise on Plane and Spherical Trigonometry, 10th edition; Lippincott, 1850; pp. 47 ff.

If $f(x) = \cos x$,
then

$$f(x+h) - f(x) = -h\sin x - \tfrac{1}{2}h^2\cos(x+\theta h).$$

When x is very small, the difference

$$\cos(x+h) - \cos x$$

is insensible and irregular.

If $f(x) = \tan x$,
then

$$f(x+h) - f(x) = h\sec^2 x + h^2\sec^2(x+\theta h)\tan(x+\theta h).$$

When x is nearly a right angle, the difference

$$\tan(x+h) - \tan x$$

is insensible and irregular.[4]

In the case of the logarithmic trigonometric functions we have the following situations:

(1) The difference, $\log\sin(x+h) - \log\sin x$ is *irregular* when the angle x is small, and it is *insensible* and *irregular* when the angle x is nearly a right angle.

(2) The difference, $\log\cos(x+h) - \log\cos x$ is *insensible* and *irregular* when the angle x is small, and *irregular* when the angle x is nearly a right angle.

(3) The difference, $\log\tan(x+h) - \log\tan x$ is *irregular* when the angle x is either very small or is nearly a right angle.

The magnitude of the term $\frac{1}{2}h^2f''(x+\theta h)$, where $0 < \theta < 1$, and the angles x and h are given in circular measures, can be considered from the point of view of the angle h and the "step-interval" of the tabulation of the trigonometric function.

1. If the tabulation is given at step-intervals of 10′, then the greatest value of $\frac{1}{2}h^2$ is

$$\frac{1}{2}\left[\frac{10\pi}{60\cdot 180}\right]^2 = \tfrac{1}{2}(0.003)^2 = 0.000005.$$

This cannot influence the last significant digit on the extreme right of a trigonometric function (whether natural or logarithmic) tabulated to four significant digits.

2. If the tabulation is given at step-intervals of 6′, then the greatest value of $\frac{1}{2}h^2$ is

---

[4] See Hobson: *op. cit.*, p. 149.

$$\tfrac{1}{2}\left[\frac{6\pi}{60 \cdot 180}\right]^2 = \tfrac{1}{2}(0.0017)^2 = 0.0000015.$$

This cannot influence the last significant digit on the extreme right of a trigonometric function (whether natural or logarithmic) tabulated to four significant digits.

3. If the tabulation is given at step-intervals of 10″, then the greatest value of $\frac{1}{2}h^2$ is

$$\tfrac{1}{2}\left[\frac{10\pi}{60 \cdot 60 \cdot 180}\right]^2 = \tfrac{1}{2}(0.00005)^2 = 0.000000001.$$

This cannot influence the last significant digit on the extreme right of a trigonometric function (whether natural or logarithmic) tabulated to seven significant digits.[5]

II. Accuracy of Tables of the Natural Trigonometric Functions

The accuracy of a table of a natural trigonometric function is determined by the relative error. In the process of the analysis that follows, we shall use angles given in circular measures. We shall denote the approximate measure of an angle by a, whose exact value is A, and whose absolute error is α, and whose relative error is $\wp_a$. Then

$$|a| \leqslant |A| + |\alpha|, \quad |\wp_a| = \frac{|\alpha|}{|A|}.$$

The relative error of a natural trigonometric function is then

$$\frac{f(|A| + |\alpha|) - f(|A|)}{f(|A|)}.$$

In the analysis that follows we shall make use of the representation of the trigonometric functions by series.

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots, \quad (x^2 < \infty),$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots, \quad (x^2 < \infty),$$

$$\tan x = x + \frac{x^3}{3} + \frac{2x^5}{15} + \frac{17x^7}{315} + \frac{62x^9}{2835} + \cdots + \\ + \frac{2^{2n}(2^{2n} - 1)B_{2n-1}x^{2n-1}}{(2n)!} + \cdots, \quad (x < \tfrac{1}{4}\pi^2)$$

---

[5] It should be noted that we do not use the term "to so many decimal places." There will be an error in the case of log sin x, given to seven significant digits, if $x < 5°$, and consequently a similar error will occur in the case of the log cos x, given to seven significant digits, if $x > 85°$. See Hobson: *op. cit.*

Where

$$\frac{B_{2n-1}}{(2n)!} = \frac{2}{(2^{2n}-1)\pi^{2n}}\left[1+\frac{1}{3^{2n}}+\frac{1}{5^{2n}}+\frac{1}{7^{2n}}+\cdots\right]$$

The B_{2n-1} are known as the Bernoulli numbers.

A. ACCURACY OF TABLES OF NATURAL SINES

The relative error of the sine of an approximate angle a is

$$\frac{\sin(|A|+|\alpha|)-\sin(|A|)}{\sin(|A|)}$$

We have then

$$\frac{\sin(|A|+|\alpha|)-\sin(|A|)}{\sin(|A|)} = \frac{\sin(A+\alpha)}{\sin A} - 1,$$

assuming hereafter that A and α are taken in their absolute values.

Let us examine the ratio

$$\frac{\sin(A+\alpha)}{\sin A}$$

by means of the representation of the sine functions of the numerator and of the denominator by series. We have then, factoring out in the numerator $(A+\alpha)$ and in the denominator A, and remembering that

$$\frac{A+\alpha}{A} = 1+\wp_{a},$$

$$(1+\wp_a)\,\frac{1-\dfrac{(A+\alpha)^2}{3!}+\dfrac{(A+\alpha)^4}{5!}-\cdots}{1-\dfrac{A^2}{3!}+\dfrac{A^4}{5!}-\cdots}$$

Performing the division of the two series in the above expression we have:

$$1-\left[\frac{2A}{3!}-\frac{4A^3}{5!}+\frac{6A^5}{7!}-\frac{8A^7}{9!}+\cdots+\frac{2(n-1)A^{2n-1}}{(2n+1)!}-\right.$$
$$\left.-\frac{2nA^{2n+1}}{(2n+3)!}+\cdots\right]\alpha + \text{terms in } \alpha^2 \text{ and in higher powers of } \alpha.$$

Write the following dominant series:

$$\frac{A}{3}+\frac{A^3}{3.2^2}+\frac{A^5}{3.2^4}+\frac{A^7}{3.2^6}+\cdots+\frac{A^{2n-1}}{3.2^{2n-2}}+\frac{A^{2n+1}}{3.2^{2n}}+\cdots$$

The absolute value of each and every term in the series within the brackets is less than the corresponding term of the dominant series. Moreover, since

$$0 \leqslant A \leqslant \frac{\pi}{2},$$

the dominant series is convergent, and its sum has

$$\frac{4A}{12 - 3A^2}$$

as its limit. The ratio test for the series within the brackets shows that

$$\left|\frac{2nA^{2n+1}}{(2n+3)!}\right| : \left|\frac{2(n-1)A^{2n-1}}{(2n+1)!}\right| = \frac{nA^2}{(2n+3)(2n+2)(n-1)} < 1,$$

which → 0 as $n \to \infty$.

Hence the series within the brackets is absolutely convergent.

We shall make use of the following theorems from the Theory of Series concerning the alternating series[6]

$$x_1 - x_2 + x_3 - x_4 + \cdot \cdot \cdot ,$$

where

$$|x_1| > |x_2| > |x_3| > |x_4| > \cdot \cdot \cdot > |x_n|,$$
$$|x_n| < \epsilon, \quad (\epsilon \to 0), \quad (n \to \infty).$$

The alternating series is convergent, and its sum S is > 0, and the remainder, after the sum of n terms is taken, is less than x_{n+1} in its absolute value. Let

$$S_{2n+1} = (x_1 - x_2) + (x_3 - x_4) + \cdot \cdot \cdot + (x_{2n-1} - x_{2n}) + x_{2n+1}.$$

Then

$$|S_{2n+1}| > |x_1 - x_2| > 0.$$

Write

$$S_{2n+1} = x_1 - (x_2 - x_3) - (x_4 - x_5) - \cdot \cdot \cdot .$$

We conclude that the sum S_{2n+1} is monotonically decreasing, and

$$|S_{2n+1}| < |x_1 - (x_2 - x_3)| < |x_1|.$$

Hence,

$$|x_1 - x_2| < \text{Limit } S_{2n+1} < |x_1 - (x_2 - x_3)|.$$

We also note that

$$S_{2n+1} = S_{2n} + x_{2n+1},$$

but, since

$$|x_{2n+1}| < \epsilon, \quad (\epsilon \to 0), \quad (n \to \infty),$$

$$\text{Limit } S_{2n} = \text{Limit } S_{2n+1}.$$

---

[6] See Pierpont, J.: Functions of a Complex Variable; Ginn, 1914; pp. 30-32.

Therefore, a limit of the sum of the series, S_m, $(m \to \infty)$, exists. The series is convergent, and

$$0 < |S| < |x_1|.$$

Consider the remaining terms, after the sum S_m is taken,

$$S'_n = x_{n+1} - x_{n+2} + x_{n+3} - x_{n+4} + \cdot \cdot \cdot .$$

This is also an alternating series, and

$$0 < |S'| < |x_{n+1}|,$$

and the number n can be chosen so that

$$|S'| < |x_{n+1}| < \epsilon, \quad (\epsilon \to 0).$$

The above results can be applied to the series within the brackets. We have then that the sum of this series is (denoting it by S)

$$\frac{A}{3} - \frac{4A^3}{5!} < S < \frac{A}{3} - \frac{4A^3}{5!} + \frac{6A^5}{7!}.$$

Since

$$0 \leqslant A \leqslant \frac{\pi}{2},$$

we shall take the upper limit of A, that is, $\frac{\pi}{2}$. We have then that

$$0.395 \leqslant S \leqslant 0.406.$$

We can safely assume that the sum of the series within the brackets is $S = 0.4$.

We have then that the relative error of the natural sine is (disregarding the terms which contain α^2 and the higher powers of α)

$$\frac{\sin (A + \alpha)}{\sin A} - 1 = (1 + \wp_a)(1 - 0.4\alpha) - 1,$$

where A and α are taken in their absolute values.

The absolute error α of an angle in the case of the tables of natural sines which are given for the intervals of the angles of or between $10'$ and $1'$, and hence the apparent error of the same angle as well, is $\leqslant 0.5'$. The maximum value of the absolute (apparent) error α in circular measure is then 0.000015. The upper limit of the relative error $\wp_a$ of the angle a (a is an approximate number) is either 0.001 or 0.0001 (in case the apparent error has a first significant digit equal to or less than the first significant digit on the extreme left of the number expressing the total number of minutes in the approximate number a).

The upper limit of the relative error of the natural sine is then either

$$(1 + 0.001)\ (1 - 0.000006) - 1,$$

or

$$(1 + 0.0001)(1 - 0.000006) - 1.$$

In either of the cases the numerical value of 0.4α does not and cannot appreciably affect the numerical value of the upper limit $\wp'_a$ of the approximate number a.

We have then that in the case when the angles are given at the intervals of or between 10′ and 1′, the relative error of the natural sine and the relative error of its corresponding angle have the same upper limit. This signifies that the tables of the natural sines which are given for angles at the intervals of or between 10′ and 1′ *should be given to four significant digits.* If the angle is less than 16°40′, three significant digits are sufficient.

The absolute error α of an angle in the case of the table of natural sines which are given for angles at the intervals of or between 10″ and 1″, and hence the apparent error of the angle as well, is $\leqslant 0.0000024$ in circular measure, or $\leqslant 0.5''$ in angular measure. The upper limit of the relative error $\wp_a$ of the angle a (a is an approximate number) is in this case either 0.00001 or 0.000001 (in cases when the apparent error has a first significant digit equal to or less than the first significant digit on the extreme left of the number expressing the total number of seconds in the approximate number, that is, angle a).

Then the upper limit of the relative error of the natural sine is either

$$(1 + 0.00001)(1 - 0.00000096) - 1,$$

or

$$(1 + 0.000001)(1 - 0.00000096) - 1.$$

In the first case the numerical value of 0.4α does not and cannot appreciably affect the numerical value of the upper limit $\wp'_a$ of the approximate number a. In the second case, however, the numerical value of 0.4α does affect the numerical value of the upper limit $\wp'_a$ of the approximate number a. It should be remembered, however, that 0.4 represents the upper limit of the sum of the series when the angle is very close to $\frac{\pi}{2}$.

We have then that in the case when the angles are given at the intervals of or between 10″ and 1″, the relative error of the natural sine and the relative error of its corresponding angle either have the same upper limit, or the upper limit of the relative error of the natural sine may be of the order "one" in comparison with the relative error of the angle itself. This happens when the apparent error of the angle has a first significant digit which is equal to or is less than the first significant digit of the extreme left of the number representing the total number of seconds in the angle. It follows then that the natural sines which are given for angles at the intervals of or between 10″ and 1″ *should be given to six significant digits.* If the angle is less than 20°46′40″, five significant digits are sufficient. The tables of the natural sines given to six significant digits will take care of the cases when the first significant digit of the apparent error of the angle is equal to or is less than the first significant digit on the extreme left of the number expressing the total number of seconds in the approximate number a (the angle) in the process when the apparent error of the natural sine is calculated (see sec. III below).

B. ACCURACY OF TABLES OF THE NATURAL COSINES

The relative error of the cosine of an approximate angle a is

$$\frac{\cos(|A| + |\alpha|) - \cos(|A|)}{\cos(|A|)}.$$

We have then

$$\frac{\cos(|A| + |\alpha|) - \cos(|A|)}{\cos(|A|)} = \frac{\cos(A + \alpha)}{\cos A} - 1,$$

assuming that A and α are taken in their absolute values. The ratio

$$\frac{\cos(A + \alpha)}{\cos A}$$

can be replaced by

$$\frac{\sin\left[\frac{\pi}{2} - (A + \alpha)\right]}{\sin\left[\frac{\pi}{2} - A\right]}.$$

In the above ratio

$$0 \leqslant A \leqslant \frac{\pi}{2}.$$

If we proceed along the same lines of analysis as in section II, A, above we can obtain the expression for the relative error of the natural cosine, which is

$$(1 + \wp_a)(1 - 0.4\alpha) - 1,$$

where $\wp_a$ is the relative error of the angle (which is given in angular measure), α is the absolute (apparent) error of this same angle a, and 0.4 is the upper limit of the sum of the series within the brackets (in sec. II, A), but in this series A is replaced by $\left(\frac{\pi}{2} - A\right)$ and the limit of A for which the series thus formed is summed is 0.

From the above we conclude that the natural cosine of a given angle a should have as many significant digits as the natural sine of this same approximate angle has. When the angles are given at the intervals of or between 10′ and 1′, the relative errors of the natural sine and of the natural cosine and the relative error of their corresponding angle have the same upper limit. This signifies that the tables of the natural cosines, which are given for angles at the intervals of or between 10′ and 1′ *should be given to four significant digits.* If the angle is less than 16° 40′ three significant digits are sufficient. When the angles are given at intervals of or between 10″ and 1″, the relative error of the natural sine and the relative error of the natural cosine and the relative error of their corresponding angle either have the same upper limits, or the upper limits of the relative errors of the natural sine and of the natural cosine may be of the order "one" in comparison with the relative error of their corresponding angle. This signifies that the tables of the natural cosines, which are given for angles at the intervals of or between 10″ and 1″ *should be given to six significant digits.* If the angle is less than 20° 46′ 40″ five significant digits are sufficient. The tables of the natural cosines given to six significant digits will take care of the cases when the first significant digit of the apparent error of the angle is equal to or is less than the first significant digit on the extreme left of the number expressing the total number of seconds in the approximate angle a in the process when the apparent error of the natural cosine is calculated (see sec. III below).

C. THE ACCURACY OF THE TABLES OF THE NATURAL TANGENTS

The relative error of the tangent of an approximate angle is

$$\frac{\tan(|A| + |\alpha|) - \tan(|A|)}{\tan(|A|)}.$$

We have then

$$\frac{\tan(|A| + |\alpha|) - \tan(|A|)}{\tan(|A|)} = \frac{\tan(A + \alpha)}{\tan A} - 1,$$

assuming that A and α are taken in their absolute values throughout.

Let us examine the ratio

$$\frac{\tan(A+\alpha)}{\tan A}$$

by means of representation of the natural tangent functions of the numerator and of the denominator by series. We have then, after $(A + \alpha)$ is factored out in the numerator, and A is factored out in the denominator, and also after the substitution

$$\frac{A+\alpha}{A} = 1 + \wp_a$$

is made, the following expression:

$$(1+\wp_a)\frac{1+\dfrac{(A+\alpha)^2}{3}+\dfrac{2(A+\alpha)^4}{15}+\dfrac{17(A+\alpha)^6}{315}+\dfrac{62(A+\alpha)^8}{2853}+\cdots}{1+\dfrac{A^2}{3}+\dfrac{2A^4}{15}+\dfrac{17A^6}{315}+\dfrac{62A^8}{2835}+\cdots}.$$

Performing the division of the two series in the above expression we have

$1 + \left[\frac{2A}{3} + \frac{14}{45}A^3 + \frac{124}{945}A^5 + \cdots\right]\alpha +$ terms in α^2 and in higher powers of α.

Applying D'Alembert's test to the series within the brackets above[7] we have that the ratio of the nth term to the $(n - 1)$st term [and because the series

$$1 + \frac{1}{3^{2n}} + \frac{1}{5^{2n}} + \frac{1}{7^{2n}} + \frac{1}{9^{2n}} + \cdots$$

in the general term of the tangent series is convergent, and its sum has a limit such that the sum S of this series[8]

$$1 < S < 1 + \epsilon, \quad (\epsilon \to 0, \quad n \to \infty)]$$

is less than 1, the series within the brackets is convergent.

If we form a geometric series with the ratio $\frac{7}{15}A^2$, and we take $A < \frac{\pi}{2}$, the sum of the series within the brackets is for all the values

$$0 \leqslant A < \frac{\pi}{2},$$

$$\frac{10A}{15 - 7A^2}.$$

---

[7] See Pierpont: *op. cit.*, p. 42.

[8] See Pierpont: *op. cit.*, p. 29.

If $A = \frac{\pi}{2} - \epsilon$, ϵ is positive and very small, the sum of the series within the brackets ranges at intervals of 10° as follows: 0.118, 0.247, 0.400, 0.602, 0.903, 1.420, 2.696, 10.318. For generality we shall take the arithmetic mean of these values, that is, 2.1. We have then for the relative error of the natural tangent the following expression:[9]

$$(1 + \wp_a)(1 + 2.1\alpha) - 1.$$

The absolute error α of the angle in the case of the tables of the natural tangents which are given for the intervals of the angles of or between $10'$ and $1'$, and the apparent error of the angle as well, is $\leqslant 0.5'$. The maximum value of the absolute (apparent) error of the angle in circular measure is then 0.000015. The upper limit of the relative error $\wp_a$ of the angle a (a is an approximate number) is either 0.001 or 0.0001 (in case the apparent error of the angle has a first significant digit equal to or less than the first significant digit on the extreme left of the number expressing the total number of the minutes in the approximate angle a).

Then the upper limit of the relative error of the natural tangent is either

$$(1 + 0.001)\,(1 + 0.00003) - 1,$$

or

$$(1 + 0.0001)\,(1 + 0.00003) - 1.$$

In the first case the numerical value of 2.1α does not and cannot appreciably affect the numerical value of the upper limit $\wp'_a$ of the relative error of the approximate number a. In the second case, however, when the first significant digit of the apparent error of the angle is equal to or is less than the first significant digit on the extreme left of the number expressing the total number of minutes in the approximate angle a, the numerical value of 2.1α does affect the numerical value of the upper limit $\wp'_a$ of the relative error of the approximate number a. (See footnote 9.)

We have then that when the angles are given at intervals of or between $10'$ and $1'$, the relative error of the natural tangent and the relative error of its corresponding angle either have the same upper limit, or the upper limit of the relative error of the natural tangent may be of the order "one" in comparison with the relative error of the angle itself. This happens when the apparent error of the ap-

[9] It should be remembered that the variation of the tangent function when the angle approaches a right angle becomes very rapid, and therefore, the generality is not affected if we use 2.1 as the sum of the series within the brackets.

proximate angle has a first significant digit which is equal to or is less than the first significant digit on the extreme left of the number expressing the total number of minutes in the angle. It follows then that the natural tangents which are given for angles at the intervals of or between 10′ and 1′ *should be given to four significant digits.* The tables of the natural tangents given to four significant digits will take care of the cases in which the first significant digit of the apparent error of the angle is equal to or is less than the first significant digit on the extreme left of the number expressing the total number of minutes in the angle in the process when the apparent error of the natural tangent is calculated (see sec. III below).

The absolute error of an angle in the case of the tables of the natural tangents which are given for the intervals of the angles of or between 10″ and 1″, and the apparent error of this same angle as well, is $\leqslant 0.0000024$ in circular measure, or $\leqslant 0.5''$ in angular measure. The upper limit of the relative error of the angle a (a is an approximate number) is in this case either 0.00001 or 0.000001 (when the apparent error of the angle has a first significant digit equal to or less than the first significant digit on the extreme left of the number expressing the total number of seconds in the approximate angle a).

Then, the upper limit of the relative error of the natural tangent is either

$$(1 + 0.00001)(1 + 0.000005) - 1,$$

or

$$(1 + 0.000001)(1 + 0.000005) - 1.$$

In either of the cases the value of 2.1α may affect appreciably the numerical value of the upper limit, $\wp'_a$, of the relative error $\wp_a$ of the approximate angle a. It should be remembered, however, that 2.1 represents the average sum of the series for a range of values of A from 0° to 80° at 10° intervals. Moreover, for natural tangents above 60° this influence of $\frac{10A}{15 - 7A^2}\alpha$ may be, and really is, very appreciable.

We have then that in the case when the angles are given at the intervals of or between 10″ and 1″, the relative error of the natural tangent and the relative error of its corresponding angle either have the same upper limit, or the upper limit of the relative error of the natural tangent may be of the order "one" in comparison with the relative error of the angle itself. It follows, then, that the natural tangents which are given for angles at the intervals of or between

$10''$ and $1''$ *should be given to six significant digits.* The tables of the natural tangents given to six significant digits will take care of the cases in which the first significant digit of the apparent error of the angle is equal to or is less than the first significant digit on the extreme left of the number expressing the total number of seconds in the approximate angle a in the process when the apparent error of the natural tangent is calculated (see sec. III below).

The same conclusions can be reached if we remember that

$$\tan a = \frac{\sin a}{\cos a}.$$

Then

$$|\wp_{(\tan a)}| \leqslant |\wp_{(\sin a)}| + |\wp_{(\cos a)}|,$$

and, if the results of section II. A and section II. B are applied, the relative error of the natural tangent is determined.

III. THE ABSOLUTE ERROR OF A TRIGONOMETRIC FUNCTION

Let a be an approximate number expressing the number of the degrees, minutes, and seconds in an angle. The absolute error of a is α, and the relative error of a is $\wp_a$, and the exact value of a is A such that

$$0 < A_2 < A < A_1.$$

Then

$$|a| \leqslant |A| + |\alpha|,$$

and

$$f(|a| \leqslant f(|A| + |\alpha|),$$

and the absolute error of a trigonometric function is

$$f(|A| + |\alpha|) - f(|A|).$$

In the case of the cosine, the cotangent, and the cosecant

$$f(|a|) \geqslant f(|A| + |\alpha|),$$

and the direction of the absolute error of $f(|a|)$ is negative.

In the analysis of the trigonometric function in section II the expression $f(|A|)$ was used without any reference to $f(|A|)$, which since it is a trigonometric function is also approximate and hence has an absolute error of its own. There is one exception, however (provided the angles are exact):

a. In the case of the natural sines:

$$\sin 0°, \sin 30°, \text{ and } \sin 90°.$$

b. In the case of the natural cosines:

$$\cos 0°, \cos 60°, \text{ and } \cos 90°.$$

c. In the case of the natural tangents:

$$\tan 0°, \tan 45°, \text{ and } \tan 90° \text{ (which is indeterminate).}$$

Let then,

$$f(|A|) = |S| + |\sigma|,$$

and

$$f(|A| + |\alpha|) \leqslant |S'| + |\sigma'|.$$

Then

$$f(|A| + |\alpha|) - f(|A|) \leqslant |S' - S| + |\sigma'| + |\sigma|.$$

This is the absolute error of a natural trigonometric function, as well as of a logarithmic trigonometric function.

If $f(|a|) = \sin(|a|)$, then[10] the absolute error portion of the natural sine is

$$|\alpha| \cos(|A|).$$

If $f(|a|) = \cos(|a|)$, then the absolute error portion of the natural cosine is

$$-|\alpha| \sin(|A|).$$

If $f(|a|) = \tan(|a|)$, then the absolute error portion of the natural tangent is

$$|\alpha| \sec^2(A).$$

The expressions for the respective apparent error portions of the trigonometric functions are similar to the above expressions for the absolute error portions.

From the above we conclude that to the absolute (apparent) error of the true value part of the trigonometric function we must add the absolute (apparent) error part which arises from the absolute (apparent) error of the angle, and the final result constitutes the absolute (apparent) error of the trigonometric function.

The determination of the absolute (apparent) error of the trigonometric function is very important, for it enables one to proceed with the relative error test and to determine the number of the significant digits in the true value portion of the natural trigonometric function.

The number of the digits in the mantissa of the logarithmic trigonometric function is determined by the absolute error and the relative error tests which are applied to the natural trigonometric functions.[11]

---

[10] See sec. 1 of this chapter.

The tables of the trigonometric functions, both natural and logarithmic, should be constructed as follows:

1. A table of the natural trigonometric functions should be based on the number of the significant digits and not on the number of decimal places in the true value of the natural trigonometric function. A numerical value of a natural trigonometric function should generally be given to the same number of significant digits as there are significant digits in the approximate angle corresponding to that function.

2. A table of a logarithmic trigonometric function should be based on the number of the significant digits in the true value portion of the corresponding natural trigonometric function; that is, the mantissas of the logarithmic trigonometric functions should have k and $(k + 1)$ digits, if the table of the corresponding natural trigonometric function is given to k significant digits.

3. To the right of every tabulation of the trigonometric function, whether natural or logarithmic, there should be a column with the apparent errors of the true value portions of the trigonometric functions.

4. Since it is impossible to tabulate apparent errors of positive and of negative directions at the same time, all the apparent errors of the trigonometric functions should be tabulated as positive, and on account of this the apparent errors of the true value portions of the trigonometric functions will be

$$10^{-k} > \text{apparent error} \geqslant 10^{-(k+1)},$$

and some apparent errors may be 0 (zero), but this should not convey the impression that the true value portion is "exact."

5. If the proportional parts or the mean differences are tabulated, the apparent error portion of the proportional parts or of the mean differences should be tabulated alongside, on the right, of each individual proportional part or mean difference.

6. Since the cosine is decreasing while the angle is increasing, and the values of the cosine functions, whether natural or logarithmic, are read upward from the bottom of the table of the sine function, the portion of the apparent error, which is obtained by means of the proportional parts, as well as the true value portion of the apparent

---

[11] See Chap. XX, sec. II.

error, should be calculated in the following manner: The true value portion of the apparent error of the cosine function is positive, whereas the proportional parts, which are to be subtracted, yield a negative portion of the apparent error. In order to render this last portion of the apparent error positive, the lower limit of the true value of the angle a, A_2, if such is given, should be changed to the upper limit A_1, and similarly the calculation of the cosine function should be performed from the upper limit A_1 of the angle a. For example, the calculation of cos (42°17′ + 0.3′) is performed as follows: We calculate from (42°18′ − 0.7′), and this is performed from (42°20′ − 2′ − 0.7′). Cos (42°17′ − 0.3′) is calculated from the angle (42°20′ − 3′ − 0.3′). Thus the proportional parts and the respective apparent errors are added. The same applies to the tables of the cotangents.

All other rules for computation with and calculation of the trigonometric functions, whether natural or logarithmic, that are given in the textbooks of trigonometry hold good for the trigonometric functions given in the tables of the type described above.

CHAPTER XXII

HISTORY OF APPROXIMATE COMPUTATION

1. General Remarks

The literature related to the original contributions to the development of the theory and the methods of Approximate Computation is very limited. Although the construction of numerical tables of the logarithmic functions was begun in the sixteenth century, and the construction of tables of numerical values of the natural trigonometric functions dates back to at least seventeen hundred years[1] before the time when the first tables of logarithms were published, and although approximate solutions of equations as well as approximate calculations of roots of numbers were topics of interest to mathematicians, there is no direct evidence that the problem of Approximate Computation and the methods involved in it ever attracted the attention of mathematicians at any date prior to the beginning of the nineteenth century.

The fact that certain quantities could be represented by numerical values which were only approximate was recognized by some of the ancient Greek mathematicians.[2] Such quantities were $\sqrt{2}$, π, etc. The theoretical development of the methods of Approximate Computation, however, is comparatively recent; probably because of the fact that the methods of Approximate Computation are based on the philosophy and on the theory of errors. The earliest work on the Theory of Errors was by Thomas Simpson (1756). The philosophy of Thomas Simpson was based on the assumption of a crude form of the Law of Normal Probability. The Law of Normal Probability was first stated by Abraham De Moivre on November 12, 1733, in his paper "Approximatio ad Summarum Terminorem Binomii $\overline{a+b}\backslash^n$ in Seriem expansi."[3]

---

[1] See Smith, D. E.: History of Mathematics; Ginn, 1925; Vol. II, Chap. VIII. Hereafter quoted as Smith, D. E.: History of Mathematics.

[2] Smith, D. E.: History of Mathematics, Vol. II, pp. 114, 254, 307–313.

[3] The translation of this paper is found in De Moivre: The Doctrine of Chances (1738), pp. 235–243. See Smith, D. E.: A Source Book in Mathematics; McGraw-Hill, 1929; pp. 566–575, in which this paper appears, edited by Professor Helen M. Walker.

II. The Notation for Approximate Numbers

The term "significant figures," that is, the term which corresponds to "significant digits," was introduced about the sixteenth century.[4] The use of this term by the early writers was directly related to the use of the Hindu-Arabic numerals, and at no time did the early writers connect the idea of significant digits with the nature of an approximate number. The term "significant digits," if not connected with the idea of approximate numbers, is rather vague, and the fact that this term was retained up to the present time can be explained by the gradual development of the theory of Approximate Computation, by the practical usefulness of numerical tables of the transcendental and other functions (logarithmic, trigonometric, etc.), and by the application of mathematics to science.

The earliest recognition of the fact that operations upon approximate numbers lead to approximate numbers was made by Johann Kepler (1571–1630). Although this was not stated by Kepler explicitly, he was the first to employ certain methods which have been in common use since. This suggests that Kepler was aware of the fact that computations with approximate numbers result necessarily in approximate numbers.

When employing the methods of the contracted multiplication (for which Kepler gives credit to Bürgi and Praetorius) he forced by one unit the digit on the extreme right of the final numerical result.[5] In other words, Kepler should be credited with the introduction of the method of rounding off digits.

Kepler employed the signs + and − in denoting the fact that rounding off the numbers led to a result in "defect" and in "excess" respectively. In the case of the excess he employed the − sign only if the error was between 0.25 and 0.5 of a unit in the last place on the right of the number.[6] In a somewhat similar manner this method was used by Johann Widman who published a textbook on commercial arithmetic[7] in which he denotes by + and − the excess and the deficiency respectively in packages in merchandise.

Of note are the following methods of writing rounded off numbers.

---

[4] Smith, D. E.: History of Mathematics, Vol. II, p. 16.

[5] Kepler, J.: Opera Omnia; ed. Frisch 7; Greifswald, 1855; p. 296.

[6] Kepler, J.: Chilias Logarithmorum; Marpurgi, 1634; Opera 7, p. 392. See Mehmke, R., in the Bibliography, Appendix 4.

[7] Widman, Johann: Behede und hubsche Rechnung auf allen kauffmanschaft; Leipzig, 1489. See Smith, D. E.: History of Mathematics, Vol. I, pp. 257–258.

If, after the rounding off, the last digit on the extreme right is 5, and this digit 5 is obtained by forcing a 4, then some writers[8] place a bar over this 5. Some writers print such a 5 in light type, whereas a 5 which is not a result of forcing a 4 is printed in bold type.[9]

In order to distinguish between an approximate number and an exact number Gernerth placed a point on the extreme right of every exact number.[10] This point served the purpose of denoting the last significant digit on the extreme right of an approximate number.

The general practice of writing the approximate numbers with their respective apparent errors was begun with the introduction of the method of least squares by Legendre.[11]

The introduction of the method of the relative errors in the processes of Approximate Computation is no doubt due to the French mathematician Vieille,[12] who in 1854 published a treatise on this subject.

III. THE "ESSAI SUR LES NOMBRES APPROXIMATIFS"

The first use of the apparent error in the analysis of the processes of Approximate Computation was made in a pamphlet by an unknown author and published in Paris in the seventh year of the French Revolution (1799) under the title:[13]

Essai sur les nombres approximatiſs
ou recherche analytique d'une méthode sûre pour obtenir
du premier coup et rigoureusement, dans le calcul de ces nombres, toute
l'exactitude qu'on desire
À Paris
chez Duprat Libraire, Quai des Augustines
An VII

or: A treatise (essay) on approximate numbers, or an analytical investigation of [illegible]e method by which to obtain (in the process) in computation at once and rigorously all the exactness that is desired.

---

[8] See Appendix 1, "Rounding Off Digits," sec. 1.

[9] See Gauss, F. G.: Fünfstellige Logarithmische Taffeln; Berlin, 1870.

[10] See Gernerth, A.: Fünfstellige Logarithmentaffeln; Wien, 1866, Einleitung.

[11] Legendre, A. M.: "Sur la méthode des moindres quarrés," in his "Nouvelles méthodes pour la détermination des orbites des comèts," Paris, 1805.

[12] Vieille, L. B. J.: Théorie générale des approximations numériques: Paris, 1852, 2nd ed. 1854.

[13] See "Anonymous" in the Bibliography, Appendix 4. The author of this pamphlet suggested the notion of the upper limit of the relative error.

The copy of this work which is found in the New York Public Library gives no information about the author, except that on the outer cover the name of Jean François Auguste Massabiau is written. It is very doubtful whether Massabiau was the author of this pamphlet, which is the earliest work on Approximate Computation known to appear in print.

The analysis of the processes of Approximate Computation in the *Essai* covers the four arithmetical operations (addition, subtraction, multiplication, and division) as well as the process of raising to a power.

After a somewhat lengthy discussion of the nature of approximate numbers the author of this pamphlet defines approximate numbers "mathematically." He distinguishes two types of numbers, that is, exact numbers and approximate numbers. The difference between the exact value and its approximation is defined as the "primitive error" (*l'erreur primitif*) which approaches a limit without ever reaching it. This primitive error is never greater than one half of a unit of the last decimal place of a number. The representation of all approximate numbers is reduced to stating them to a certain number of decimal places. The primitive error can be either positive or negative. If a number is correct to n decimal places, then its primitive error (positive) is

$$\frac{1}{10^n} - \frac{5}{10^{n+1}} - \frac{1}{\infty}.$$

The symbol $\frac{1}{\infty}$ stands for an "infinitesimal" in the modern sense.

Disregarding $\frac{1}{\infty}$ we have

$$\frac{1}{10^n} - \frac{5}{10^{n+1}} = \frac{1}{2 \cdot 10^n}.$$

The author of this pamphlet almost grasped the idea of the relative error. He defined *the degree of the primitive error* as follows: Let S be the number of a measure (an approximate number), and $\frac{1}{2 \cdot 10^n}$ be the primitive error. Then

$$S \cdot \frac{1}{2 \cdot 10^n} = \frac{S}{2 \cdot 10^n}.$$

Let

$$S = a10^x + b10^{x-1} + \cdot \cdot \cdot + t10^0,$$

where

$$a, b, c, \cdot \cdot \cdot t, < 10.$$

Let

$$\frac{1}{2 \cdot 10^{n-x}}$$

be a general expression of the primitive error. Then

$$\frac{S}{2 \cdot 10^{n}} = \frac{a10^{z} + b10^{z-1} + \cdot \cdot \cdot + t10^{0}}{2 \cdot 10^{n}} < \frac{1}{2 \cdot 10^{n-x}}$$

and

$$a10^{z} + b10^{z-1} + \cdot \cdot \cdot + t10^{0} < 10^{z},$$

and[14]

$$z = x + 1.$$

The development of the rules for the various operations was obtained by the procedure that has been in use since that time. We will illustrate the case of subtraction. Let

N and N' be the approximate numbers,
Q and Q' be the corresponding exact numbers,
e and e' be the corresponding primitive errors,
$Q - Q'$ be the difference of the exact numbers,
$N - N'$ be the difference of the approximate numbers.

Then

$$Q = N \pm e, \text{ and } Q' = N' \pm e'.$$

The primitive error of the difference of the two approximate numbers is

$$(Q - Q') - (N - N'),$$

or

$$(N \pm e) - (N' \pm e') - N + N' = \pm e \mp e'.$$

IV. GAUSS'S THEORIA MOTUS

The problem of Approximate Computation attracted the attention of Gauss. In his work on the motion of the heavenly bodies[15] he devoted several sections to the methods of Approximate Computation. The views of Gauss are very interesting, because Gauss was the first to suggest that the errors (the apparent errors) should be "affected by the same signs," that is, that apparent errors should be of the same direction when the processes of Approximate Computation take place. This suggestion was probably overlooked by all the subsequent writers on the subject of Approximate Computation. Below are the two sections from the Theoria Motus[16] in which Gauss's philosophy of Approximate Computation is stated.

---

[14] See Chap. XIV, sec. XIII.

[15] Gauss, C., Fr.: Theoria Motus Corporum Coelestium, Hamburg, 1809.

Section 30

Since none of the numbers which we take out from logarithmic and trigonometrical tables admit of absolute precision, but are all to a certain extent approximate only, the results of all calculations performed by the aid of these numbers can only be approximately true. In most cases indeed the common tables, which are exact to the seventh place of decimals, that is, never deviate from truth either in excess or in defect beyond half a unit in the seventh figure, furnish more than the required accuracy, so that the unavoidable errors are evidently of no consequence; nevertheless it may happen that in special cases the effect of the errors of the tables is so augmented that we may be obliged to reject a method, otherwise the best, and substitute another in its place. Cases of this kind can occur in the computations which we have just explained; on which account it will not be foreign to our purpose to introduce here some inquiries concerning the degree of precision allowed in these computations by the common tables. Although this is not the place for a thorough examination of this subject, which is of the greatest importance to the practical computer, yet we will conduct the investigation sufficiently far for our own subject, from which point it may be further perfected and extended to other operations by any one requiring it.

Section 31

Any logarithm, sine, tangent, etc., whatever (or in general any irrational quantity whatever taken from the tables) is liable to an error which may amount to a half a unit in the last figure; we designate this limit of error by ω, which therefore is in the common tables = 0.00000005. If now the logarithm, etc., cannot be taken directly from the tables, but must be obtained by means of interpolation, this error may be slightly increased from two causes. *In the first place*, it is usual to take for the proportional part, when (regarding the last figure as unity) it is not an integer, the next greatest or least integer; and in this way, it is easily perceived, this error may be increased to just within twice its actual amount. But we shall pay no attention to this augmentation of the error, since there is no objection to our affixing one more than another decimal figure to the proportional part, and it is very evident that, if the proportional part is exact, the interpolated logarithm is not liable to a greater error than the logarithms given directly in the tables, so far indeed as we are authorized to consider the changes in the latter as uniform.[17] Thence arises *another* increase of the error, that this

---

[16] Taken from the English translation of the Theoria Motus by Ch. H. Davis. Boston, 1857.

[17] Gauss was not aware of the fact that the apparent error of a logarithm is directly related to the relative error of the corresponding approximate number. There is no evidence in any of the works of Gauss that he knew of the properties of the relative error. The computations with logarithms necessarily involve the use of the relative error. See Chapter XX.

> last assumption is not rigorously true; but this also we pretermit, because the effect of the second and higher differences (especially where superior tables computed by Taylor are used for trigonometrical functions) is evidently of no importance and may readily be taken into account if it should happen to turn out a little too great. In all cases, therefore, we will put the maximum unavoidable error of the tables = ω, assuming that the argument (that is, the number the logarithm of which, or the angle the sine, etc., of which, is sought) is given with the strict accuracy. But if the argument itself is only approximately known, and the variation ω' of the logarithm, etc. (which may be defined by the method of differentials[18]) is supposed to correspond to the greatest error to which it is liable, then the maximum error of the logarithm, computed by means of the tables, can amount to $\omega + \omega'$.
>
> Inversely, if the argument corresponding to a given logarithm is computed by the help of the tables, the greatest error is equal to that change in the argument which corresponds to the variation ω in the logarithm, if the latter is correctly given, or to that which corresponds to the variation $\omega + \omega'$ in the logarithm, if the logarithm can be erroneous to the extent of ω'. It will hardly be necessary to remark that ω and ω' must be affected by the same sign.[19]
>
> If several quantities, correct within certain limits only, are added together, the greatest error of the sum will be equal to the sum of the greatest individual errors affected by the same sign; wherefore, in the subtraction also of quantities approximately correct the greatest error of the difference will be equal to the sum of the greatest individual errors. In the multiplication or division of a quantity not strictly correct the maximum error is increased or diminished in the same ratio as the quantity itself.[20]

Whether or not Gauss knew of the *Essai sur les nombres approximatifs* is difficult to determine. It can be safely assumed that [illegible] might have developed his theory of Approximate Computation independently. However, his work in the field of Approximate Computation is very superficial. He was very little interested in the subject, never beyond the immediate needs of his practical computations.

The most interesting contribution of Gauss to the theory of Approximate Computation, besides the recognition of the fact that the direction of the errors of the approximate numbers involved in the computation processes should be the same, is the use of the process of differentiating for the purpose of obtaining the expressions for the apparent errors of formulas.[21] A subsequent writer in the field of Approximate Computation who elaborated on this method was Guyou.[22]

---

[18] Gauss obtained the expressions for the apparent errors by differentiating the formulas used in the processes of the computation.

[19] This passing remark of Gauss is of great importance in the processes of Approximate Computation because the proper procedure hinges on it.

[20] This last statement should have led Gauss to the discovery of the properties of the relative error.

V. THE LAW OF ERRORS

The Law of Errors should properly be called the Normal Law of Errors. An extensive analysis of the properties of this law was made by Laplace and Gauss, although a proof of the law had already been given by De Moivre in 1733.[23] Laplace and Gauss proved the law independently of each other, attacking the problem from entirely different angles.

Gauss[24] derived the Law of Errors by postulating the following: "When any number of equally good direct observations M, M', M'', . . . of an unknown magnitude x are given, the most probable is their *arithmetic mean*." This is known as the Postulate of the Arithmetic Mean. This postulate is not universally true, nor is a postulate of a "most probable value" true of the arithmetic mean only. For some physical observations the arithmetic mean is not the most probable value. There are also a great number of other functions of which it can be said that "as the number of the observations increases indefinitely, the function $f(x)$ tends to the true value of x."[25]

Having assumed this Postulate of the Arithmetic Mean, Gauss deduced his Normal Law of Errors

$$f(x) = \frac{h}{\sqrt{2\pi}} e^{-\frac{1}{2}h^2x^2},$$

where h is a certain constant, and $f(x)$ is the relative frequency of an error.

Laplace deduced the Normal Law of Errors from the following considerations:[26] he derived the sum of the errors of a large number of observations, the errors of which may be equally $-n$, $-(n-1)$,

---

[21] See sec. 32 of the Theoria Motus.

[22] Guyou, E.: Note sur les approximations numériques. Paris, 1909. Also: Nouv. Annals de Mathèmatique (3), 8. 1889, p. 165.

[23] See sec. I of this chapter.

[24] Theoria Motus, sec. 177.

[25] See: Whittaker, E. T. and Robinson, G.: The Calculus of Observations; Blackie & Son, Ltd., London, 1926; pp. 215–219.

[26] Laplace, P. S.: Oeuvres Completes, Vol. VII. Paris, 1886. Book 1, sec. 36, and Book 2, secs. 18 and 19. A translation of the portion of Book 2 is found in Smith, D. E.: A Source Book in Mathematics, pp. 588–595.

$-(n-2), \cdots, -2, -1, 0, 1, 2, \cdots, (n-2), (n-1), n$, that is, the probability of each error is $\frac{1}{2n+1}$. He then derived the expression for the probability that the sum of these errors will be between the limits $+h$ and $-h$. This probability is (we shall use the modern notation, now in common use)

$$P(h) = \theta(H) = \frac{2}{\sqrt{\pi}} \int_0^H e^{-u^2} du,$$

where

$$H = \frac{\sqrt{6}}{2\sqrt{n(n+1)s}}\, h,$$

s is the number of the observations,

and

n is the maximum (in absolute value) error.[27]

It is indeed of considerable interest that Laplace recognized the fact that at times the added errors may all be affected by the same sign.[28] He found that the sum of such errors is at the most equal to one-half of the largest possible sum of the limits of these errors. This result was overlooked by many of the subsequent writers on the theory of Approximate Computation. On the other hand, it should not have led to the so-called theory of the "compensating property" of errors in and of the measurements performed on different objects which is erroneously based on the Law of Errors.

VI. THE INTRODUCTION OF THE RELATIVE ERROR IN THE PROCESSES OF APPROXIMATE COMPUTATION

It is difficult to determine the origin of the method of the relative errors in the processes of Approximate Computation. No doubt the problem proposed by F. Wolff but never solved by him[29], "To perform a computation with the least number of digits (significant?) so that the numerical result will contain the predetermined number of digits" (literally: "be of the predetermined approximativeness") was the first problem in Approximate Computation that led to the thorough analysis of the properties of approximate numbers.

---

[27] See Kryloff, A. N.: Lekzii o priblizhonnykh vychislenijakh (Lectures on Approximate Computations), 2d ed. Leningrad. The Academy of Sciences of U. S. S. R., 1933, pp. 195–204.

[28] Laplace: *op. cit.*, book 2, sec. 19. Smith, D. E.: A Source Book in Mathematics, pp. 593–595.

[29] "Ein Rechnung mit möglichst geringem Aufwand an Ziffern so auszuführen dass die für das Ergebnis vorgeschriebene Genauigkeit gesichert ist," Wolff, F.: Theoretisch-praktische Zahlenlehre. Berlin, 1832.

In his work on Approximate Computation Vieille[30] introduced the use of the relative error, and since then many of the subsequent writers based their work on Approximate Computation on either the apparent error analysis or on the relative error criterion. Fundamentally, the properties of the relative error can be illustrated by the following example:[31]

If

$$f = \frac{ab^m \sqrt[n]{c} \cdots}{de^p \sqrt[q]{g} \cdots},$$

then, denoting the apparent errors of the numbers $f, a, b, c, d, e, g,$ . . . by $\Delta f, \Delta a, \Delta b, \Delta c,$ . . . respectively (some of the numbers may be exact) we have the relative error

$$\frac{\Delta f}{f} = \frac{\Delta a}{a} + m\frac{\Delta b}{b} + \frac{1}{n}\frac{\Delta c}{c} + \frac{\Delta d}{d} + p\frac{\Delta e}{e} + \frac{1}{q}\frac{\Delta g}{g} + \cdots.$$

The introduction of the relative error test, and the use of the apparent error (but never the combined use of these two criteria for the approximativeness) naturally led to the examination of the significant digits approximativeness of the numerical results of performed computations. Very conflicting rules have been stated from time to time by many writers. Griess,[32] for example, stated the following rules: "The product (or quotient) of two numbers: a containing m significant digits, and b containing n significant digits, $m < n$, is correct to $(m - 2)$ significant digits. If the first digit on the extreme left of a is 1 (in the text: "if a begins with a 1") then the numerical result is correct to $(m - 3)$ significant digits."

VII. THE ANALYSIS OF CONTRACTED MULTIPLICATION AND DIVISION

Contracted multiplication and division have been in common use since the sixteenth century.[33] The theoretical analysis of the problem of contracted multiplication and division is comparatively recent. Its purpose was to examine the limits of the errors of the contracted product and of the contracted quotient of two approximate numbers. In the two illustrations following we shall use the symbolism employed in this work.

---

[30] Vieille, Z. B. J.: Théorie générale des approximations numériques. Paris, 1852. 2d ed. 1854.

[31] Griess, S.: Approximations numériques. Paris, 1898, p. 18.

[32] Griess: *op. cit.*, p. 18.

[33] See sec. II of this chapter.

A. THE LIMIT OF THE ERROR OF THE CONTRACTED PRODUCT OF TWO APPROXIMATE NUMBERS[34]

Let A and B be two positive exact numbers. Their approximations are a and b respectively, and their respective absolute errors are α and β.[35] Then

$$A = a + \alpha, \quad B = b + \beta, \tag{1}$$

Let a contain m significant digits, and let b contain n significant digits, then

$$a = a_0 10^p + a_1 10^{p-1} + \cdots + a_{m-1} 10^{p-m+1}, \quad \left(p \gtreqless 0\right), \tag{2}$$

$$b = b_0 10^q + b_1 10^{q-1} + \cdots + b_{n-1} 10^{q-n+1}, \left(q \gtreqless 0\right), \tag{3}$$

and

$$|\alpha| < 0.5 \cdot 10^{p-m+1}, \quad |\beta| < 0.5 \cdot 10^{q-n+1}. \tag{4}$$

Then, from (1) we have

$$AB = (a + \alpha)(b + \beta) = ab + a\beta + b\alpha + \alpha\beta,$$

and from (2), (3), and (4) we have

$$a|\beta| < 0.5 \cdot 10^{p+q-m-n+2}(a_0 10^{m-1} + a_1 10^{m-2} + \cdots + a_{m-1}), \tag{5}$$

$$b|\alpha| < 0.5 \cdot 10^{p+q-m-n+2}(b_0 10^{n-1} + b_1 10^{n-2} + \cdots + b_{n-1}). \tag{6}$$

If

$$(b_0 10^{n-1} + b_1 10^{n-2} + \cdots + b_{n-1}) > (a_0 10^{m-1} + a_1 10^{m-2} + \cdots + a_{m-1})$$

then

$$b > a,$$

and the right side of (6) is greater than the right side of (5). Then there is no need to carry the product beyond the units in the 10^k $(k = p + q - m + 1)$, place of significance.

The contracted product $P = ab$ is then obtained in the following manner. The least accurate number is taken as the multiplicand. The partial product $ab_0 10^q$ has its last digit on the extreme right in the 10^k place of significance.

---

[34] Stolz, O. und Gmeiner, J. A.: Theoretische Arithmetik. Leipzig, 1902, p. 180.

[35] Stolz and Gmeiner worked with the apparent errors.

The absolute error of the product P is obtained from the formula

$$AB - P = (a + \alpha)B - P = (aB - P) + B\alpha.$$

The $B|\alpha|$ part has a limit which is identical with the limit of $b|\alpha|$ in (6), that is, $0.5(b_0 + 1)10^k$.

If $m < n$, then $|aB - P| < 1.5m10^k$,
if $m = n$, then $|aB - P| < [1.5m + 0.5(a_0 - 1)]10^k$.

The last follows from the evaluation of the partial products

$$ab_0 10^q, \quad ab_1 10^{q-1}, \cdots, \text{(when } n > 2).$$

If $\beta = 0$, i.e. $B = b$, then, if $m = n$,

$$|ab - P| < (1.5m - 1)10^k.$$

B. THE LIMIT OF THE ERROR OF THE CONTRACTED QUOTIENT OF TWO APPROXIMATE NUMBERS[36]

From (1) above we have

$$\frac{A}{B} = \frac{a + \alpha}{b + \beta} = \frac{a}{b} + \left[\frac{\alpha}{B} - \frac{a\beta}{Bb}\right]. \tag{7}$$

The limits for

$$\frac{|\alpha|}{B} \quad \text{and} \quad \frac{a|\beta|}{Bb}$$

are obtained from the limits for $|\alpha|$ and $|\beta|$ in (4). The greater limit determines the place of the significance to which the quotient should be contracted. Let this place of significance be in the 10^k units. Denote by Q the quotient, and by R the remainder of the division. The quotient is then contracted to the units of the order 10^k. Evaluate the product $P = Qb$ (b the multiplier) which has its last significant digit on the extreme right in the 10^{q+k} place of significance. We have then

$$a = P + R, \quad [|R| \leqslant (0.5b_0 + 1)10^{q-k}], \tag{8}$$

Let

$$Qb = P + F. \tag{9}$$

Then

$$\frac{A}{B} - Q = \left[\frac{A}{B} - \frac{a}{b}\right] + \left[\frac{a}{b} - Q\right]$$

The first part (in parentheses) on the right of the equality sign above is evaluated by means of the formula (7). For the second part we have from (8) and (9)

---

[36] Hočevar, F.: Lehrbuch der Arithmetik und Algebra. Wien, 1901, p. 79.

$$\frac{a}{b} - Q = \frac{a - bQ}{b} = \frac{R - F}{b}.$$

The limits of the error F and of the number Qb are evaluated by the method illustrated in *a* above.

The two methods described are lengthy and laborious. If the two authors here quoted had resorted to the combined use of the apparent error and the relative error tests, much work would have been eliminated, and the entire analysis would have been simplified.

APPENDIX

APPENDIX I

ROUNDING OFF DIGITS

1. General Remarks

Rounding off digits serves a twofold purpose. By performing this operation the order of *the precision* and the order of *the accuracy* of the numerical value of a performed or given measurement and represented by an approximate number is changed. In the processes of Approximate Computation the change of the order of precision or of the order of accuracy of the numerical values of approximate measures is very often necessary.

The rules usually given for rounding off the digits are as follows:

1. If a number, given to a certain number of significant digits, is to be rounded off to a stated number of significant digits, the digits that are to be dropped should be replaced by zeros. In the case when the digits that are to be dropped are located to the right of the decimal point, the use of the zeros is not necessary.

2. If the first digit on the left of those that are to be dropped is 5, 6, 7, 8, or 9, and this digit is to be dropped also, then the first digit on the extreme right of the number, which is to be retained, should be increased by a unity (1). This process is known as "forcing the digit."

3. If after the forcing, the significant digit on the extreme right is 5, a *bar* should be placed over it in order to indicate that if there should be a necessity to drop this 5, the next digit on its left should remain unchanged. For example: 3,464,832 is rounded off to 3,46$\bar{5}$,000, and this is rounded off to 3,460,000.

Some writers present the following rule: When the first digit on the left that is to be dropped is 5, then the first significant digit of the number on the extreme right, which is retained, should be forced if it is odd, and should be left unchanged if it is even.[1]

---

[1] This rule is entirely arbitrary, for it is just as logical and just as consistent to force the digit when it is even and to leave it unchanged when it is odd, as to force it when it is odd and to leave it unchanged when it is even. This rule is based on the assumption that in the case of a large number of computations with approximate numbers the increases in the digits are compensated by those numbers that are left unchanged. Its effectiveness, since errors in and of measurements of different objects cannot compensate one another, is doubtful.

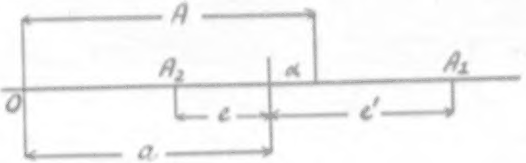
II. THE NATURE OF AN APPROXIMATE NUMBER

Since approximate numbers are represented by the true value portions and by their corresponding apparent errors (or at times with the relative errors), the process of rounding off the digits must be considered from the point of view of a measurement. An approximate number represents the numerical value of a measurement, and this value is located within two limits, known as the upper limit and the lower limit of the true value of the approximate number representing the numerical value of the measurement. The expression "limit" used here is synonymous with the term "bound."

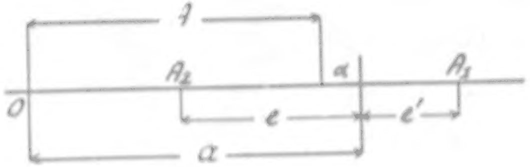
Whether the approximate number is represented by the upper limit of the true value and its corresponding apparent error, or represented by the lower limit of the true value and its corresponding apparent error, the magnitude of the approximate number is the same in either case. In other words, an approximate number can be expressed in two forms which essentially represent the same magnitude of a measurement. The two forms, that is, one with the upper limit of the true value, and the other with the lower limit of the true value, represent the same measurement, but in different senses:

a. The *lower limit* of the true value corresponds to a measurement taken *in defect*; and
b. The *upper limit* of the true value corresponds to a measurement taken *in excess*.

This can be seen from the following diagram.



or



From the above we have that

$$a = A_2 + e = A_1 - e',$$

and if an approximate number is correct to k significant digits, then

$$|e| + |e'| = 10^{p+n-k+1},$$

where n is an integer > 0; p is an integer $\gtreqless 0$; k is an integer > 0; and e and e' are the apparent errors.

III. Rounding Off Approximate Numbers

The use of the upper limit or of the lower limit of the true values of an approximate number in the process of Approximate Computation is restricted only by the rules for Approximate Computation.

Since it is assumed that the apparent error of an approximate number[2] is

$$\leqslant 0.5 \cdot 10^{p+n-k+1},$$

where k is the number of the significant digits in the true value portion of the approximate number, the distance between the lower and the upper limit of the true value of an approximate number is then

$$10^{p+n-k+1},$$

that is, the distance is one unit of the value in the last place of significance on the extreme right of the true value portion of an approximate number.[3] This distance represents the sum of the absolute values of the apparent errors of an approximate number represented once with the lower limit of the true value, and once with the upper limit of the true value. For example:

$$a = 6.43 + 0.003,$$
$$a = 6.44 - 0.007,$$
$$0.003 + 0.007 = 0.01.$$

The above illustrates the principles that are involved in rounding off approximate numbers. In addition to rules 1, 2, and 3 of section I, we have the following rules:

1. If an approximate number is given as correct to m significant digits, and it is required to round it off to k significant digits ($k < m$), the magnitude of the apparent error of the rounded off approximate

---

[2] See Chap. XIV, sec. XI.

[3] See Chap. XIV, sec. XI.

number will be one or several units in the $(k+1)$st place of significance, counting from the left to the right.

2. The digit in the $(k+1)$st place of significance should be rounded off according to rules 1, 2, and 3 of section I.

3. The rounded off approximate number should have an apparent error always

$$\leqq 0.5 \cdot 10^{p+n-k+1}.$$

That is, if the resultant apparent error has a first significant digit >5, the limit of the true value portion of the rounded off approximate number should be changed. This will result in an apparent error of the above form.

Examples:

1. Round off $14{,}567 + 0.3$ to 3 significant digits. We have then

$$14{,}500 + 70, \quad \text{or} \quad 14{,}600 - 30.$$

2. Round off $643.5447 - 0.00004$ to 5 significant digits. We have then

$$643.54 + 0.00\bar{5}.$$

3. Round off $643.54 + 0.00\bar{5}$ to 4 significant digits. We have then

$$643.5 + 0.045, \quad \text{or} \quad 643.5 + 0.0\bar{5}.$$

The method of rounding off the digits by changing the limits as illustrated above leaves the order of precision of the rounded approximate number unchanged.

IV. THE RELATIVE ERROR

The relative error of an approximate number depends on the magnitude of the apparent error of the same approximate number, and the direction of the relative error of an approximate number is the same as the direction of its corresponding apparent error.

Whenever rounding off of an approximate number is effected or a change in the limits of the true value of an approximate number takes place, the relative error of the changed approximate number should be calculated accordingly. It should be remembered that in the case of the change of the limits of the true value of an approximate number (unless the apparent error has a first significant digit on the left $= 5$) the new relative error of the approximate number may be of one order higher (or one order lower) than the relative error of the approximate number in the original state.

V. ROUNDING OFF APPROXIMATE NUMBERS WHOSE ERRORS HAVE THE SIGN ±

The process of rounding off approximate numbers whose apparent errors have the sign ±, deserves special attention. When an approximate number is given with an apparent error of this type, the number representation assumes that the exact value of the approximate number is located between the upper and the lower limits of the true value, that is,

$$A_2 < A < A_1.$$

These limits of the true value control the precision of an approximate number, and if a number represents a measurement or a result of a performed computation, once it is of a certain type and order of precision this precision *cannot* be improved. It may however be lowered; that is, the apparent error of the approximate number may, at will, be increased.

We shall state without proof the following three theorems which are of importance in the process of rounding off approximate numbers whose apparent errors are with the signs ±.

1. If the upper limit of the absolute error of an approximate number is increased by substituting for it a number greater than the given upper limit of the absolute error, then the interval within which the exact value of the approximate number lies is widened.

2. If an approximate number a is such that

$$A_2 < a < A_1,$$

then for the upper limit of the absolute error we can take the greatest of the difference $|a - A_2|$ or $|A_1 - a|$.

3. If we take for A the value

$$\frac{A_1 + A_2}{2},$$

and for α_1 the value

$$\frac{A_1 - A_2}{2},$$

then the numerical value of α_1 *is at its minimum.* For examples in rounding off numbers of the type discussed in this section consult Chapters XI and XII.

APPENDIX 2

NOTE ON ANTILOGARITHMS

LET the table of the antilogarithms be given for the logarithms at the intervals of the arguments of $10^{-(k-1)}$, that is, the logarithms are given with $(k-1)$ places in their mantissas, and the kth place is to be interpolated. For two consecutive tabulations in a table of antilogarithms we have

$$M \log_e N$$

and

$$M \log_e N + 10^{-(k-1)},$$

where M is the *modulus*: $M = 0.43 + 0.00\bar{5}$.

Let

$$M \log_e N + 10^{-(k-1)} = M \log_e (N+n).$$

Interpolating for some value of the argument between the values of $M \log_e N$ and $M \log_e N + 10^{-(k-1)}$, that is, for

$$M \log_e N + t10^{-k} = M \log_e (N+x),$$

$$1 \leqslant t \leqslant 9, \quad x < n,$$

we have:

$$\frac{M \log_e N + t10^{-k}}{M} = \log_e (N+x),$$

and

$$N + x = e^{\log_e N + \frac{t10^{-k}}{M}} = Ne^{\frac{t10^{-k}}{M}}.$$

Then

$$x = N(e^{\frac{t10^{-k}}{M}} - 1).$$

Expand the expression within the parentheses in power series. We have then

$$x = N\left\{\frac{t10^{-k}}{M} + \frac{1}{2}\left[\frac{t10^{-k}}{M}\right]^2 + \cdots\right\};$$

the terms in higher powers than the second can be disregarded, for, as it will be shown, the generality of the conclusions will not be affected by this.

In order that the principle of proportional parts may hold in the case of antilogarithms, the terms within the parentheses, beginning with the term in the second power, must not affect the last significant digit on the extreme right of the true value portion of the approximate number N. Since the mantissa of the logarithm contains k digits, the approximate number N should have either k or $(k + 1)$ significant digits.[1] Evaluate the term in the second power within the parentheses. We have then:

$$N2.65(t10^{-k})^2 < N\tfrac{1}{2}\left[\frac{t10^{-k}}{M}\right]^2 < N2.7(t10^{-k})^2.$$

Assume for the time being that $t = 1$, and let us examine the effect produced by

$$N\tfrac{1}{2}\,\frac{10^{-2k}}{M^2} \tag{A}$$

on the last significant digit on the extreme right of the approximate number N given as correct to either k or $(k + 1)$ significant digits. Denote the first significant digit on the extreme left of the approximate number N by x_1.

Let N have a true value portion with k significant digits. Then, if $x_1 < 4$, the first significant digit on the extreme left of (A) will appear in the $(k + 1)$st place, and this will not affect the last significant digit on the extreme right of the true value portion of N. If $x_1 \geqslant 4$, the first significant digit on the extreme left of (A) will appear in the kth place, and this will affect the last significant digit on the extreme right of the true value portion of N.

Let N have a true value portion given with $(k + 1)$ significant digits. If $x_1 < 4$, then the first significant digit on the extreme left of (A) will appear in the kth place, and this will affect the last significant digit on the extreme right of the true value portion of N. If $x_1 \geqslant 4$, then the first significant digit on the extreme left of (A) will appear in the $(k - 1)$st place, and this will affect the last two significant digits on the extreme right of the true value portion of N.

If $t \geqslant 2$, the true value portion of N will be always affected.

The above shows that the method of proportional parts does not generally hold in the case of antilogarithms. Linear interpolation cannot be applied to antilogarithms in general.

---

[1] See Chap. XX, Secs. II and IV.

APPENDIX 3

SYMBOLS USED IN THE TEXT

$=$	The "equality" sign.	$a = b$; a equals b.
$\neq$	The "inequality" sign.	$a \neq b$; a is not equal to b.
$\equiv$	The "equivalence" sign.	$a \equiv b$; a is equivalent to b.
$> <$	The inequality signs.	$a < b$; a is less than b.
		$a > b$; a is greater than b.
		$a \geqslant b$; a is equal to or is greater than b.
		$a \leqslant b$; a is equal to or is less than b.
		$a \overset{>}{\underset{<}{=}} b$; a is equal to, greater than, or less than b.

$\approx$	This sign denotes approximate equality.
$\underset{\sim}{\equiv}$	This sign is to be read "belongs to the order."
$\sim$	This sign implies that $a \sim b$, a approaches b as closely as possible, but there is always a difference between these two quantities, however small this difference may be. In other words, this symbol reads: "a approaches b asymptotically."
$\lvert a \rvert$	The absolute value of a. Thus $\lvert -5 \rvert = 5$.
$\epsilon \longrightarrow 0$	This means that ϵ is an infinitesimal, and it approaches zero as a limit. (Chapter XIII.)

$k = 0, 1, 2, \cdots, n$ means that k assumes the values of $k = 0$, $k = 1$, $k = 2$, $\cdots$, $k = n$, each value once only.

$x_k = 0, 1, 2, \cdots, 9$; $k = 0, 1, 2, \cdots, n$ means that x_k assumes the values of the first ten integers, 0, 1, 2, $\cdots$, 9, and the value of each integer may be repeated. For example, $x_2 = 5$ and $x_7 = 5$. At the same time, however, k assumes the values from zero to n once only.

$\sum$	The summation sign	$\sum_{k=0}^{n} a_k = a_0 + a_1 + a_2 + \cdots + a_n.$
$\prod$	The product sign	$\prod_{i=0}^{n} a_i = a_0 a_1 a_2 \cdots a_n.$

$\underset{a \to b}{\text{Limit}}\ x$ reads: "The limit of x as a approaches b."

The following theorems from algebra are used and referred to in the text:

1. $$|a + b| \leqslant |a| + |b|.$$

2. $$|a - b| \leqslant |a| + |b|.$$

3. If

$$\frac{a}{A} = \frac{b}{B} = \frac{c}{C} = \cdot\ \cdot\ \cdot = \frac{n}{N},$$

then

$$\frac{a + b + c + \cdot\ \cdot\ \cdot + n}{A + B + C + \cdot\ \cdot\ \cdot + N} = \frac{a}{A}.$$

4. If

$$0 < \frac{a}{A} < 1 \quad \text{and} \quad 0 < \frac{b}{B} < 1,$$

and

$$\frac{a}{A} \gtreqless \frac{b}{B},$$

then

$$0 < \frac{ab}{AB} < 1,$$

and

$$\frac{ab}{AB} < \frac{a}{A}, \quad \frac{ab}{AB} < \frac{b}{B}.$$

5. If

$$0 < \frac{a}{A} < \frac{b}{B} \leqslant \frac{c}{C} \leqslant \cdot\ \cdot\ \cdot < \frac{n}{N},$$

then

$$\frac{a}{A} < \frac{a + b + c + \cdot\ \cdot\ \cdot + n}{A + B + C + \cdot\ \cdot\ \cdot + N} < \frac{n}{N}.$$

APPENDIX 4

BIBLIOGRAPHY

THE LITERATURE ON APPROXIMATE COMPUTATION

Abramoff, N. M. Tekhnicheskije vychislenija (Technical computations). Moscow, 1928.

Adhémar, R. See Montessus, R. et Adhémar, R.

Airy, G. B. Theory of errors of observations. Cambridge, 1861.

Anonymous. Essai sur les nombres approximatifs, ou recherche analytique d'une méthode sûre pour obtenir du premier coup et rigoureusement, dans le calcul *de ces nombres*, toute l'exactitude qu'on desire. À Paris, chez Duprat Libraire. An VII (1799).

Aurelius, N. F. De gradu praecisionis, quam tabulae logarithmicae admittunt. Upsaliae, 1813.

Babb, M. J. Some remarks on approximate computation. The Mathematics Teacher, 3, 1910–1911, p. 43. New York.

Bennett, A. A., *Milne, W. E.*, and *Bateman, H.* Numerical integration of differential equations. Report of Committee on numerical integration, Division of Physical Sciences, National Research Council. Bulletin 92 of The National Research Council. Washington, D. C., 1933. (Contains an extensive bibliography on the Theory of Approximations.)

Besikovitsch, J. Priblizhonnyje vychislenija (Approximate computations). Moscow, 1932.

Biermann, O. Vorlesungen über mathematische Näherungsmethoden. Braunschweig, 1905.

Bourget, J. Théorie élémentaire des approximations numériques. Paris, 1860.

Bovier-Lapierre, G. Traité élémentaire des approximations numériques. Lyons, 1863.

Bradis, V. Teoria i praktika vychislenij (The theory and practice of computations). Moscow, 1932.

Kak nado vychisliatj priblizhonnyje vychislenija (How to perform approximate computations)—in three parts. Moscow, 1932.

Umnozhenije priblizhonnykh tchisel (Multiplication of approximate numbers). Bulletin of the Physico-Mathematical Society of Kazan University. Vol. XXV, ser. II, 1925.

Pravila podstchota tzifr (The rules for counting the digits). Matematicheskoje obrasovanije (Mathematical Education). Nos. 1 (pp. 40–44) and 2 (pp. 71–82). Moscow, 1928.

Opyt obosnovanija niekotorykh prakticheskikh pravil dejstvij nad priblizhonnymi tchislami (An experiment for the purpose of the determination of certain practical rules for operations with approximate

These pages were
~~THIS PAGE WAS~~ MISSING FROM THE DOCUMENT THAT WAS SUBMITTED TO ERIC DOCUMENT REPRODUCTION SERVICE:

pg. 285 and 286

Van der Flit, A. P. Arithmetika priblizhonnykh vychislenij (The arithmetic of approximate computations). Prague, 1922.

Vieille, L. B. J. Théorie générale des approximations numériques. Paris, 1854.

Volberg, O. A. Vlijanije okruglenija na raspredelenije pogreshnostej (The influence of the rounding off on the distribution of errors). Berichte des pädagogischen Institutes in Twer. Heft 5, 1929, pp. 119–152.

Walker, H. M. and *Sanford, V.* The accuracy of computation with approximate numbers. The Annals of Mathematical Statistics. March 1934. Ann Arbor.

Werkmeister, T. Praktisches Zahlenrechnen. Berlin, 1921.

Wolff, F. Theoretisch-praktisches Zahlenlehre. Berlin, 1832.

Xavier, A. Théorie des approximations numériques et de calcul abrégé. Paris, 1909.